THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Flip-chip Integrated Superconducting Quantum Processors

HANG-XI LI

Quantum Technology Laboratory
Department of Microtechnology and Nanoscience (MC2)
Chalmers University of Technology

Göteborg, Sweden 2025

Flip-chip Integrated Superconducting Quantum Processors
HANG-XI LI
ISBN 978-91-8103-153-9

© HANG-XI LI, 2025

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr. 5611
ISSN 0346-718X

Quantum Technology Laboratory
Department of Microtechnology and Nanoscience (MC2)
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
A flip-chip integrated superconducting quantum processor sitting inside a printed circuit board

Chalmers Digitaltryck
Göteborg, Sweden 2025

Flip-chip Integrated Superconducting Quantum Processors
HANG-XI LI
Quantum Technology Laboratory
Department of Microtechnology and Nanoscience (MC2)
Chalmers University of Technology

Abstract

On the path toward fault-tolerant quantum computing—an endeavor motivated by the prospect of solving otherwise intractable computational problems in fields such as quantum chemistry, materials, and optimization—a key challenge is to scale up the number of quantum bits of information (qubits) a quantum computer can host while not degrading their performance. To this end, the superconducting quantum processor (SQP) has its advantages due to its flexible design, compatibility with microchip manufacturing processes, and addressability by microwaves generated by commercially available equipment.

This thesis is a demonstration of the scalability of SQPs. By adopting 3-dimensional integration technologies used in semiconductor manufacturing, flip-chip integrated SQPs can host dozens to hundreds of qubits, compared to the smaller number of qubits a single-chip architecture can accommodate. The first part of this thesis shows how we transferred the design of individual components of the SQP—qubits, couplers, readout resonators, and Purcell filters—into a flip-chip architecture while maintaining good qubit coherence and high control-and-measurement performance with additional fabrication processes. We pay special attention to the interchip spacing, an additional design parameter introduced in the flip-chip architecture, which has a large influence on the parameter predictability and performance of the SQP.

The second part of the thesis shows how we used these individual components to design a scaled-up SQP. The design workflow of a multi-qubit SQP, from parameter design to layout, is elaborated in detail. This workflow has resulted in a 25-qubit flip-chip integrated SQP, without degrading the qubit coherence and gate performance, further demonstrating the scalability of flip-chip integrated SQPs. We speed up this design workflow by introducing an analytic design method for superconducting resonators based on conformal mapping techniques, which we use to design readout resonators with parameters that are not affected by variations of the interchip spacing.

Keywords: quantum computing, superconducting quantum processor, 3D integration, flip-chip integration, design workflow, simulation and design speed-up.

List of publications

This thesis is based on the work contained in the following papers:

Paper A S. Kosen, **H.-X. Li**, M. Rommel, D. Shiri, C. Warren, L. Grönberg, J. Salonen, T. Abad, J. Biznárová, M. Caputo, L. Chen, K. Grigoras, G. Johansson, A. F. Kockum, C. Križan, D. P. Lozano, G. J. Norris, A. Osman, J. Fernández-Pendás, A. Ronzani, A. F. Roudsari, S. Simbierowicz, G. Tancredi, A. Wallraff, C. Eichler, J. Govenius, and J. Bylander, "Building blocks of a flip-chip integrated superconducting quantum processor", Quantum Science and Technology **7**, 035018 (2022)

Paper B K. Grigoras, N. Yurttagul, J.-P. Kaikkonen, E. T. Mannila, P. Eskelinen, D. P. Lozano, **H.-X. Li**, M. Rommel, D. Shiri, N. Tiencken, S. Simbierowicz, A. Ronzani, J. Hatinen, D. Datta, V. Vesterinen, L. Gronberg, J. Biznarova, A. F. Roudsari, S. Kosen, A. Osman, M. Prunnila, J. Hassel, J. Bylander, and J. Govenius, "Qubit-compatible substrates with superconducting through-silicon vias", IEEE Transactions on Quantum Engineering **3**, 5100310 (2022)

Paper C L. Chen, **H.-X. Li**, Y. Lu, C. W. Warren, C. J. Križan, S. Kosen, M. Rommel, S. Ahmed, A. Osman, J. Biznárová, A. Fadavi Roudsari, B. Lienhard, M. Caputo, K. Grigoras, L. Grönberg, J. Govenius, A. F. Kockum, P. Delsing, J. Bylander, and G. Tancredi, "Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier", npj Quantum Information **9**, 26 (2023)

Paper D **H.-X. Li**, D. Shiri, S. Kosen, M. Rommel, L. Chayanun, A. Nylander, R. Rehammar, G. Tancredi, M. Caputo, K. Grigoras, L. Grönberg, J. Govenius, and J. Bylander, "Experimentally Verified, Fast Analytic, and Numerical Design of Superconducting Resonators in Flip-Chip Architectures", IEEE Transactions on Quantum Engineering **4**, 3101312 (2023)

Paper E S. Kosen, **H.-X. Li**, M. Rommel, R. Rehammar, M. Caputo, L. Grönberg, J. Fernández-Pendás, A. F. Kockum, J. Biznárová, L. Chen, C. Križan, A. Nylander, A. Osman, A. F. Roudsari, D. Shiri, G. Tancredi, J. Govenius, and J. Bylander, "Signal crosstalk in a flip-chip quantum processor", PRX Quantum **5**, 030350 (2024)

Other papers that are outside the scope of this thesis:

Paper I

Y. Lu, M. Kudra, T. Hillmann, J. Yang, **H.-X. Li**, F. Quijandría, and P. Delsing, "Resolving Fock states near the Kerr-free point of a superconducting resonator", npj Quantum Information **9**, 114 (2023)

Paper II

L. Chen, S. P. Fors, Z. Yan, A. Ali, T. Abad, A. Osman, E. Moschandreou, B. Lienhard, S. Kosen, **H.-X. Li**, D. Shiri, T. Liu, S. Hill, A.-A. Amin, R. Rehammar, M. Dahiya, A. Nylander, M. Rommel, A. F. Roudsari, M. Caputo, G. Leif, J. Govenius, M. Dobsicek, M. F. Giannelli, A. F. Kockum, J. Bylander, and G. Tancredi, "Fast unconditional reset and leakage reduction in fixed-frequency transmon qubits", arXiv, 2409.16748 (2024)

Contents

Abstract i

List of publications iii

Contents v

Acknowledgements ix

Abbreviations xi

I Thesis 1

1 Introduction 3
- 1.1 Why quantum computation? 3
- 1.2 Superconducting quantum processor (SQP) 4
- 1.3 Scaling up is necessary 5
- 1.4 Challenges ahead 6
- 1.5 Thesis outline 7

2 3-dimensional (3D) integration technology 9
- 2.1 Advanced packaging in semiconductor industry 9
- 2.2 3D integration in SQP 11
  - 2.2.1 Separating the qubits from their control lines 11
  - 2.2.2 Controlling the qubits vertically 12
  - 2.2.3 Differences from classical advanced packaging technologies 13
  - 2.2.4 Implementation requirements 14
- 2.3 Discussion of the future trend 14

3 Flip-chip integrated SQPs 17
- 3.1 3D structures in flip-chip integrated SQPs 17
- 3.2 Fabrication process 19
- 3.3 Individual qubit performance 20
- 3.4 Two-qubit pair as a unit cell 22
- 3.5 Purcell filter for qubit readout 23
- 3.6 Interchip spacing variation effect 27
  - 3.6.1 Global variation 27
  - 3.6.2 Local variation 29
- 3.7 Further developments in flip-chip architecture 30

4 Design of a multi-qubit, flip-chip integrated SQP 31
- 4.1 General design workflow 31

4.2 Architecture design . . . 32
4.3 Parameter design . . . 33
4.3.1 Constraints on the unit cell level . . . 33
4.3.2 Constraints on the chip level . . . 34
4.3.3 Parameter optimization . . . 34
4.4 Layout design . . . 35
4.4.1 Layout drawing workflow of an SQP . . . 38
4.5 Design verification . . . 39
4.6 Performance characterization after fabrication . . . 40
4.7 Discussion on scaling up SQPs . . . 42

5 Fast design of coplanar-waveguide (CPW) resonators in flip-chip architecture 45
5.1 Analytical calculation . . . 45
5.2 Comparison to 3D simulations . . . 49
5.2.1 Resonant frequency f_r . . . 49
5.2.2 Coupling quality factor Q_c . . . 49
5.3 Discussion on computational resources consumption . . . 50
5.4 Making f_r insensitive to interchip spacing . . . 52
5.5 Q-factor of the resonator's cross-section in flip-chip architecture . . . 53

6 Summary and contributions 55
6.1 Summary . . . 55
6.2 Contributions . . . 56

A Packaging of an SQP 57

B Derivation of L_l^g, C_l of a CPW line 61
B.1 Planar CPW line . . . 62
B.2 Flip-chip CPW line . . . 63
B.2.1 Top chip with the ground plane . . . 65
B.2.2 Top chip without the ground plane . . . 67

References 69

II Appended Papers 75

Paper A 77

Paper B 107

Paper C 119

Paper D 133

Paper E 147

To Ke Zou, my lovely wife.

Acknowledgements

When I look back at the journey I have walked through as a PhD student, I would say, ‘incredible’. There are many moments that are so fulfilling, and also moments that are so stressed. Most importantly, I have met so many people that I genuinely want to give my thanks for helping me along the way.

First, I would like to thank my supervisor, Jonas Bylander, for your kind support during my time at Chalmers. I could not have come to this point without your guidance on my work and help with arranging experimental resources. I greatly appreciate your detailed remarks on my writings, which always force me to think deeper about the meaning that I have written down.

I would also give my sincere thanks to Daryoush Shiri, who is not only my co-supervisor, but also my best friend in Sweden. You gave me a lot of advice, both in research and in life. It was a delightful experience to work with you, chat with you, and share my thoughts with you. I am also very grateful for your critical comments on my writings and talks, after which they always result in a much better quality.

I would also like to thank my examiner, Per Delsing, for your valuable input on my research and your kind help throughout my PhD study, and our current head of the division, Witlef Wieczorek, for your management of QTL’s research activities and support of healthy working environment.

To Anita Fadavi Roudsari and Daniel Perez Lozano, I am still very thankful for your patience in teaching me lots of design and measurement techniques at the beginning of my PhD. And Sandoko Kosen, working closely with you is a pleasant and rewarding experience. I would not have accomplished so much work without our collaboration. You are my role model for being hard-working.

My weak point is always at fabrication. Thank you a lot, Marcus Rommel, for your patient explanation and for showing me how to fabricate the chips out, one step by another. Moreover, without your effort, our flip-chip devices would not have been so good and could not have been fabricated so successfully.

I am also very thankful to the people in QTL that I always have discussions with: Giovanna Tancredi, Robert Rehammar and Andreas Nylander, as senior people in the group, the discussions with you had enriched my knowledge; and my fellow PhDs, Liangyu Chen, Christopher Warren, Janka Biznárová, Christian Križan, Amr Osman, Hampus Renberg Nilsson, Martí Gutierrez Latorre, Yong Lu, Jiaying Yang, Anuj Aggarwal, Emil Rehnman, and Achintya Paradkar, the discussions with you were always fruitful. Especially Liangyu, it was a great time when we had intense discussions during our collaboration on the qubit readout experiments. Moreover, I had a good time working with Lert Chayanun, Miroslav Dobsicek, Mamta Mamta, Irshad Ahmad, and all the current and previous members of QTL.

I should also thank many other people who have helped me along the way: Linda Brånell, for your kindness in helping me deal with administration paperwork and Swedish migration system; Lars Jönsson, for your magical making of our sample boxes; Tuure Rantanen, Kestutis Grigoras, Leif Grönberg, Marco Caputo, Joonas Govenius and other people in VTT, for completing the fabrication of flip-chip devices; Henrik Frederiksen, for your help on fabrication recipes; Graham Norris, for your early help on chip design;

Alexey Zadorozhko, for your maintenance of cryogenic fridges; Basudha Srivastava, for your teaching of quantum error correction; and Jorge Fernández-Pendás, Tahereh Abad, Anton Frisk Kockum, Göran Johansson and many people in AQP for your theoretical input at various aspects. There are definitely more people who I have discussed with and learned from, thank you very much.

To my parents, Xu-Xun Li and Yi-Qing Xie, thank you for understanding me not being able to visit you often during the past five years and for always supporting my pursuit of being a scientist.

And finally, to my love, Ke Zou, this year is our 10-year anniversary since we were together. Throughout my life, I can say my happiest moment was our wedding ceremony three years ago. Without you, I cannot imagine what my life in Sweden would look like. It is you who are always beside me, sharing my happiness, comforting my frustration, and giving me the energy to move on. Thank you so much for being here with me. This thesis is dedicated to you.

Hang-Xi Li, Göteborg, January 2025

Abbreviations

Here we list the abbreviations in the order that they first occur in the thesis.

RSA	Rivest–Shamir–Adleman scheme
SQP	superconducting quantum processor
L–C	inductor–capacitor
3D/2.5D/2D	3/2.5/2-dimensional
CPW	coplanar-waveguide
SoC	system-on-chip
I/O	input/output
PCB	printed circuit board
IC	integrated circuit
TSV	through-silicon via
CMOS	complementary metal–oxide–semiconductor
C-chip	control chip
Q-chip	qubit chip
SiP	system-in-a-package
EM	electromagnetic
SEM	scanning electron microscope
UBM	under-bump metalization
Al	aluminum
NbN	niobium nitride
In	indium
TiN	titanium nitride
TLS	two-level system
E-field	electric field
SQUID	superconducting quantum interference device
RB	randomized benchmarking
iRB	interleaved randomized benchmarking
RDL	redistribution layer
MCM	multi-chip module
SNR	signal-to-noise ratio
EDA	electronic design automation
DRC	design rule checking
CPU	central processing unit

Part I

Thesis

CHAPTER I

Introduction

When Richard Feynman in his 1982 landmark speech [1] said that 'you can simulate this (quantum mechanics) with a quantum system, with quantum computer elements,' he may not have thought that after only 40 years, quantum computers would have become so real in the world and evolve so fast. Just before the publication of this thesis, Google announced having used a 105-qubit superconducting quantum processor to demonstrate an exponential suppression of the logical qubit error rate with quantum error-correcting codes [2]. At the same time, IBM has announced having built a 1000-qubit superconducting quantum processor [3]. Both companies have used flip-chip integration technology to scale up their processors [4, 5], which will be the main topic of this thesis.

1.1 Why quantum computation?

Why do we want to have a quantum computer, a machine to realize *quantum computation*?

The basic building block of quantum computers, which gives them their unique 'quantum' nature, is the 'quantum bit' of information, or *qubit*. Unlike the state of a classical bit, which can only be in a digital 0 or 1, the state of a qubit, $|\psi\rangle$, can be in a quantum superposition of the two [6]:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle. \tag{1.1}$$

The coefficients α, β are complex numbers that must satisfy the constraint $|\alpha|^2 + |\beta|^2 = 1$. What's more, two qubits can be 'entangled' together as a joint state $|\Phi\rangle$ that cannot be separated into two single-qubit states:

$$|\Phi\rangle = c_1|00\rangle + c_2|01\rangle + c_3|10\rangle + c_4|11\rangle \neq (\alpha_1|0\rangle + \beta_1|1\rangle)(\alpha_2|0\rangle + \beta_2|1\rangle), \tag{1.2}$$

where the coefficients of this joint state satisfy $|c_1|^2 + |c_2|^2 + |c_3|^2 + |c_4|^2 = 1$.

A quantum algorithm is realized by a sequence of quantum gates operating on such joint states. Different from two-bit Boolean logic gates in 'classical' algorithms, which operate

on one two-bit state at a time [7], a two-qubit quantum gate operates on the coefficients of all computational basis states ($|00\rangle$, $|01\rangle$, etc.) of the joint state *simultaneously* and creates entanglement between qubits. This is the *quantum parallelism*, which can give a cleverly devised quantum algorithm an exponential advantage over a classical algorithm when applied to some computational problems. Here 'exponential advantage' means that the number of operations (and hence time to solution) grows as an exponential in the size of the problem (number of bits needed) in the classical algorithm, whereas it can grow much more slowly, as a polynomial, in a quantum algorithm [6].

The most prominent example of this is the quantum algorithm that Peter Shor had developed for factoring of integer numbers [8], with important applications in decryption of widely used public-key cryptography schemes such as the Rivest–Shamir–Adleman scheme (RSA) [9], which is based on the belief that factoring of integers is 'hard,' i.e., it cannot be done in polynomial time.

After the publication of Shor's algorithm, more and more efforts have been put into searching for quantum algorithms that may have an advantage over classical algorithms. Grover's algorithm [10] was shown to have quadratic speedup for unstructured database search. Simulation methods for quantum systems, just as Feynman envisioned, were proposed to try to utilize the same quantum power that Shor had discovered [11].

1.2 Superconducting quantum processor (SQP)

Among the many physical platforms that have been suggested for the realization of quantum computers [12], the superconducting quantum processor (SQP), based on circuit quantum electrodynamics [13], represents one of the most developed and promising technologies [14]. Its design flexibility and fabrication compatibility with mainstream semiconductor fabrication equipment have resulted in its rapid growth of scale [3, 15] after its first prototype was realized [16].

The qubits within an SQP are based on the non-linear behavior of the Josephson tunnel junctions [17]. Following years of evolution, a rich design landscape of superconducting qubits has emerged [7, 13], and researchers are developing SQPs based on various kinds of qubit designs [4, 18–20].

The superconducting qubit modality used in this thesis is called transmon, essentially an inductor–capacitor (L–C) oscillator whose spectrum has been rendered slightly anharmonic by the Josephson nonlinear inductance. Its name comes from its original realization using a transmission line [21]. The transmon has a large shunt capacitor, compared to its Josephson junction's own capacitance. Figure 1.1(a) shows a typical design of the transmon within the SQP. Additional control and readout components are used to manipulate the transmons and interrogate their states. A dedicated coupler (in our case, another qubit) mediates the interaction between qubits within the SQP, bringing about quantum entanglement.

Figure 1.1(b) shows a simple 2-qubit SQP design that contains all the elements just mentioned. Such a design is generally patterned on a superconducting thin film (here aluminum) on top of a dielectric substrate (silicon).

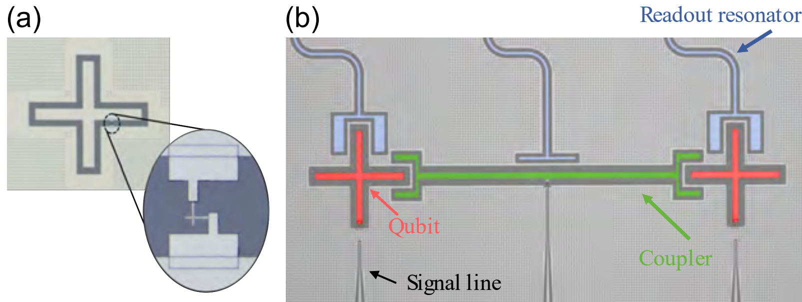


Figure 1.1: (a) Standalone transmon qubit in which the Josephson junction (zoom-in photo) connects the qubit island and the ground plane. (b) A small SQP made at Chalmers, containing two qubits connected by a coupler, along with their control and readout components. Picture adapted from Reference [22].

1.3 Scaling up is necessary

Why do we first design a small SQP like in Figure 1.1(b)? It is quite easy to see that such a 2-qubit pair design can be viewed as the unit cell of a larger SQP with a square qubit lattice. An SQP with only 2 qubits can be used to do some interesting scientific experiments [22, 23], but to demonstrate that a quantum computer is superior to its classical counterpart, thousands or even millions of qubits need to be linked together [24].

Specifically, there are two factors that result in the need for a large number of qubits. The first one comes from the necessity of using a quantum computer. For example, as Feynman discussed in his paper [1], when a quantum system is small, we can still simulate it using classical computers. Only when the system has too many variables can the benefit of using the quantum computer surpass the cost of building it. What's more, we believe that the larger the quantum system we can simulate, the more interesting phenomena we can observe, as P. W. Anderson has claimed [25]. A similar argument can also be used for all other quantum algorithms including Shor's algorithm. We always pursue those quantum algorithms that can be run *exponentially* faster than their classical counterparts for a fixed size of the problem. Conversely, you can say that when the problem is small, it may not be cost-efficient to solve it using a quantum computer.

Another factor comes from the imperfection of qubits. Perturbation from their surrounding environment is unavoidable since the qubits are not completely isolated. We describe different kinds of perturbation as different noise channels of the qubit. To overcome these noises, such that the qubit can stay at a target state for as long as we want, we can either make efforts to suppress these noise channels by better fabrication [26] and engineering [27], or to correct the errors incurred by the noise by implementing a quantum error-correcting code [6]. Similarly to classical error-correcting codes [28], a quantum error-correcting code requires redundancy, meaning that one uses more than one qubit to represent a single logical qubit in an abstract way. For example, with the

quantum repetition code, we can represent a logical qubit by N physical qubits with

$$|\psi\rangle_L = \alpha_L|0\rangle_L + \beta_L|1\rangle_L = \alpha_L|\underbrace{0...0}_{\text{N}}\rangle + \beta_L|\underbrace{1...1}_{\text{N}}\rangle. \tag{1.3}$$

Notice that $|\psi\rangle_L$ is an entangled state if α_L or β_L is not equal to zero. The more physical qubits you link together for one logical qubit, the longer the state of this logical qubit can last against errors [2, 29]. As the size of the target problem keeps increasing, better qubits are always needed. For quantum error-correcting codes, a better logical qubit means more qubits, i.e., the SQP needs to scale up.

1.4 Challenges ahead

However, scaling up an SQP is not easy. At every step of producing a working SQP, i.e., its design, fabrication, and operation, there are limits on the number of qubits that can be involved.

Typically, the finite amount of resources imposes most of such limits on the possible scale of the final produced SQP. A relatively fixed resource is needed to design a single qubit, including its control and readout components, and linked couplers. For example, each qubit's parameters need to be optimized along with surrounding qubits and then realized with the help of simulations. Therefore, trying to reduce the design cost per qubit is a must to scale up the SQP. When the qubit number is large, a small reduction in design cost per qubit can significantly affect the SQP's total resource expense.

The fabrication of an SQP not only constitutes a fixed resource cost per qubit, but also has a large overhead that increases with the scale of the SQP. Usually, the footprint of a qubit on the SQP is a fixed quantity: as the number of qubits increases, so does the overall size of the SQP. Considering the cost of fabrication, such an increase in the SQP's size likely reduces the yield of each fabrication run due to fabrication variations and defects [15], which necessitates costly process development and upgrades to the tool set.

The effort to ensure the constant operation of an SQP also should not be underestimated. As the size of the computational problem increases, so does the number of qubits involved and the complexity of the qubit control system. Continuous calibration of every qubit is needed to achieve stable high-fidelity qubit control [4]. To correctly and continuously execute quantum algorithms on the SQP, a complex software stack with layers of abstraction is built upon the physical device and needs to be maintained [30, 31]. Although similar concepts apply in classical computing, the particularities of the quantum hardware and ways of operating a quantum computer bring a whole new set of challenges to hardware and software engineering. From the hardware perspective, to enable the control and readout of every qubit within the SQP, the number of electronic control devices should align with the scale of the SQP, which results in a correspondingly increasing heat dissipation, putting high demands on the cooling system at both ambient and cryogenic environment [32].

Apart from finite resources, another challenge that scales together with the SQP is qubit interference. With an increase in size, the total crosstalk induced by other qubits

for each qubit within the SQP will also increase. The level of crosstalk must decrease rapidly with the qubit-qubit distance at a threshold rate in order to ensure each qubit's functionality, as we discuss in Paper E. This requirement brings additional architecture, design, and fabrication improvement as the SQP scales up.

Finally, as the physical size of the SQP increases with the qubit number, those noise channels that scale with chip size will transition from being negligible to significant. For example, the impact of high-energy radiation only becomes noticeable when the SQP is large enough, and the frequency of occurrence of such an impact can reach once per 10 seconds [33], which is detrimental to logical qubit state computation using quantum error-correcting codes. Eventually, much more work needs to be done to mitigate such an impact [27].

1.5 Thesis outline

As there are so many challenges ahead, scaling up a superconducting quantum processor is a complex and difficult task. This thesis covers the works that I have contributed to this project, focusing on tackling the design and simulation challenges.

Chapter 2 briefly introduces the background of 3-dimensional (3D) integration technology. First, it reviews the advanced packaging technologies used by the semiconductor industry. And then it discusses the reasons why we would like to adopt these technologies into the SQP to scale it up further, and the differences in the implementation compared to the semiconductor industry. It also discusses the requirements for implementing 3D integration technologies in SQPs.

Within the new architecture, modification of the SQP design and fabrication process is necessary. **Chapter 3** shows the architecture of the flip-chip integrated SQPs, and the modified design of every component within the SQP. In particular, we demonstrate the similar performance of these components to that of single-chip designs. It also discusses the variation effect of interchip spacing on the SQP design, an additional variable imposed by the flip-chip architecture. This chapter covers the works in Paper A, B and C.

With successful designs of SQP components in the flip-chip architecture, **Chapter 4** then discusses how to design a scaled-up flip-chip integrated SQP in detail. It goes through a design workflow that includes architecture, parameter, and layout design, and then discusses the design verification step that will become more and more necessary when the size of the SQP becomes larger. This chapter summarizes our design experiences of the SQPs used in Paper D and E.

Chapter 5 focuses on how to speed up the design and simulation of the readout resonators, a necessary component to read out the state of qubits. It shows that we can use only the 2-dimensional (2D) cross-section of the resonator to predict its resonant frequency and coupling quality factor to its readout line. Such work is important because each qubit needs one readout resonator, and when the scale of SQP is large, saving the computational resources spent on these resonators, e.g., reducing the simulation time by three orders of magnitude, can greatly accelerate the design workflow. It also shows a modified resonator design that can overcome the effect of variations of the interchip spacing discussed in Chapter 3. This chapter is based on the work of Paper D.

This thesis ends with a summary in **Chapter 6** and appended papers that have been mentioned above. Though a single PhD thesis definitely cannot contain all the aspects of scaling up an SQP, I hope this thesis can give the reader a clear overview and some understanding of the challenges we face and the solutions we propose, so as to finally build a quantum computer that can benefit society.

CHAPTER II

3-dimensional (3D) integration technology

In the history of the semiconductor industry, the chip packaging technology generally evolves together with the scaling of the chip [34]. Recently, rather than making a monolithic system-on-chip (SoC), building up a SoC by so-called chiplets [35] has become a trend to satisfy end users' growing needs of system performance while at the same time keeping the cost down, since not every functionality within the SoC needs the most advanced fabrication process. As a result, advanced packaging technologies [36] that serve to integrate these chiplets into one SoC are becoming a hot topic within the semiconductor industry. Coincidentally, in the field of superconducting quantum computing, the scale of SQPs has reached a level where adopting 3D integration technologies from advanced packaging technologies has become a necessity.

This chapter starts with a brief review of the advanced packaging technologies currently used by the semiconductor industry. We then shift our attention to quantum chips, discussing the reason why people introduce 3D integration technologies into SQPs, and the differences between classical chips and SQPs regarding these technologies. We then talk about the challenges of implementing 3D integration technologies into SQPs and these techonologies' future trends.

2.1 Advanced packaging in semiconductor industry

Following the exponential increase of the transistor number that Moore's law [37] had successfully predicted, the packaging technology has also evolved to keep up with the pace of the increase of the chips' input/output (I/O) density and data speed [34]. Following years of development, more and more complex packaging technologies have been developed to integrate the most advanced chips of each generation.

In the early times, packaging was a step after the chip was made, during which the chip is encapsulated by the epoxy molding compound and placed on top of an organic or ceramic printed-circuit board (PCB), like surface-mount technology [38]. However,

as the chip becomes denser and smaller, the first packaging step that a silicon chip experiences is more and more likely to be another silicon chip for signal routing, before being encapsulated and placed on the PCB. Such advanced packaging step often occurs in 2.5-dimensional integrated circuits (2.5D-IC) or 3D-IC [36].

A 2.5D-IC architecture typically includes an interposer made of silicon to host the connections between chiplets that have different functionalities. As shown in Figure 2.1(a), the chiplets are attached to the interposer through micro-bumps using flip-chip bonding technology. After that, the interposer is again attached to the package substrate or PCB using a different bonding technology. To reach the transistors on the top chips, signal and power go through bumps between the layers and vias within the interposer. For 3D-IC, the difference is that the connection between chips is not through an interposer, but by stacking one chip on top of another chip through flip-chip bonding, with bumps between the two chips and vias within the bottom chip to deliver signal and power to the top chip, as shown in Figure 2.1(b).

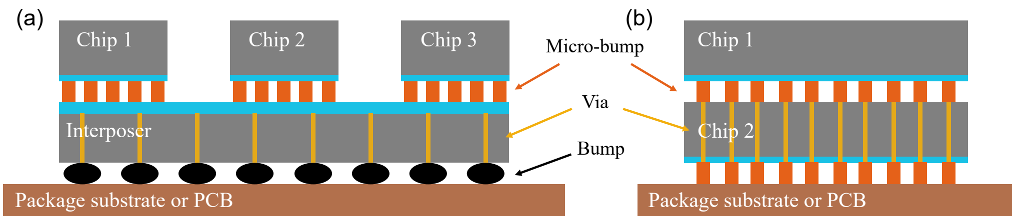


Figure 2.1: (a) 2.5D-IC architecture. (b) 3D-IC architecture.

As we can see, both 2.5D and 3D-IC have used the flip-chip bonding technology to integrate the chips together. Actually, flip-chip bonding technology was invented quite early. Historically, in order to increase the I/O port number between the chip and the PCB, people transitioned from putting the I/O ports on the four edges of the chip to the front side of the chip, such that the port number can increase with the square of the chip's size, rather than just linearly [39]. As a result, the chip then needs to be flipped and bonded using bumps to facilitate its connection to the PCB. Following the increase of the chip's transistor count and density, both the pitch between bumps and the size of the bump shrink considerably to match the increase of I/O number, but such a packaging method is always called flip-chip bonding technology.

Via connections is another technology that is always required by 2.5D and 3D-IC architecture. If the interposer or the bottom chip is made of silicon, through-silicon vias (TSVs) that are filled with conductive metal are used to realize the signal and power connection between the top side and the bottom side of the interposer or the bottom chip. When the chip's transistor density increases, the pitch of the TSVs will decrease accordingly.

2.2 3D integration in SQP

Typically, qubits within an SQP are fabricated on the surface of a silicon chip, the same substrate used for complementary metal-oxide-semiconductor (CMOS) transistors [40]. In the CMOS process, the transistor's source, drain, and gate are made within several layers on the chip surface. However, unlike the semiconductor chip with transistors that are able to proceed to build additional layers of interconnects on top of the transistors to link them together on the same chip, the general practice in SQPs is to keep qubits facing the air, so as to avoid additional losses [41] that are detrimental to the qubit coherence. As a result, the control and readout lines to address these qubits have to route on the same plane as the qubits from the edge of the SQP, and then connect to the PCB via wirebonds, as illustrated by Figure 2.2.

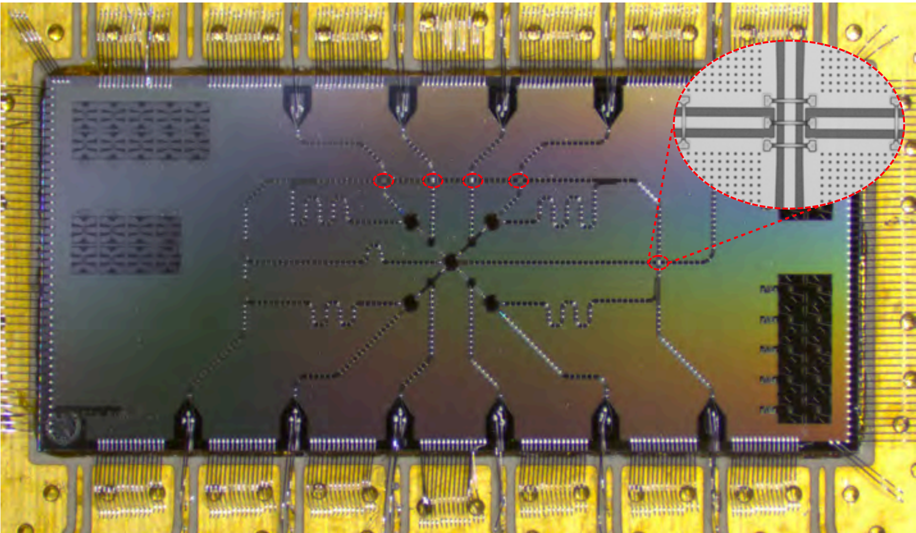


Figure 2.2: A single-chip 5-qubit SQP at our lab. There are wirebonds connecting SQP's control and readout lines from SQP's launch pads on the edges to the CPW lines on the PCB (gold). Within the SQP, the control lines on the top part of the SQP are using airbridges [42] to jump across the readout lines of the qubits (red circles with a zoom-in image of the airbridges at the crossing point.). Photo of the SQP is taken by Dr. Christopher Warren.

2.2.1 Separating the qubits from their control lines

When the SQP is scaling up, meaning the qubits are linked by couplers to form a square lattice, routing lines on the same layer as that of qubits cause problems. Since each qubit needs at least one control line, to address the center qubit of a 3×3 square lattice, crossover is unavoidable between the control line and the couplers along the way. When the qubit number is small, the crossover can be resolved by using airbridges to make the control lines jump over the coupler [43], as shown in Figure 2.2. However, as the number

of qubits increases, so do the number and the density of these crossovers. Continuing using airbridges can result in a low yield of the SQP fabrication and considerable crosstalk between couplers.

To avoid using airbridges, and at the same time satisfy the constraint that qubits should face the air first, putting all the lines on another chip's surface and making this chip's surface hover above the qubits and couplers with a distance, is a natural solution. We call the chip that hosts lines the control chip (C-chip), while the chip that hosts qubits and couplers, the qubit chip (Q-chip). The separation between the C-chip and the Q-chip can be realized by the flip-chip bonding technology similar to the technology in advanced packaging [44], with bumps located around the qubits and couplers on the Q-chip side, and the lines on the C-chip side passing through the rows of bumps, as illustrated by Figure 2.3.

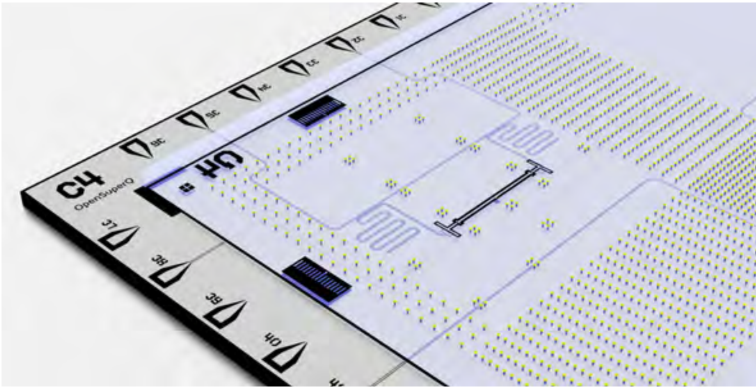


Figure 2.3: 3D model of a flip-chip 2-qubit SQP, where C-chip and Q-chip are separated by bumps sitting between the two under-bump metalization layers (gold color dots). A detailed description of this flip-chip architecture is in Chapter 3.

2.2.2 Controlling the qubits vertically

As the number of lines to control the qubits increases with the scale of the SQP, we will face the same problem as the classical semiconductor chip, in which the chip size is mostly determined by the number of I/O ports on the edge. A mere separation of the SQP into two chips cannot solve this problem. However, if the control lines can vertically approach the qubit, not only can we eliminate the crossovers, but we can also vastly reduce the required chip size, since now the allowable number of I/O ports is proportionally increased with the chip area, rather than the chip's edge length.

For single-chip SQPs, the lines can be realized by vertical pins directly penetrating from the sample holder's lid [45, 46]. If the SQP has been separated by C-chip and Q-chip using flip-chip bonding technology, the lines can then be vertically routed through the C-chip by vias, and fanned out on the back side of the C-chip, as illustrated by Figure 2.4.

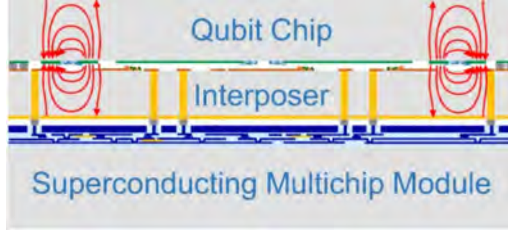


Figure 2.4: Cross-section of a flip-chip integrated SQP implemented with TSVs, in which the C-chip is called interposer, and there is a third chip under the C-chip to redistribute the lines to the outside. Figure adapted from Reference [47].

2.2.3 Differences from classical advanced packaging technologies

It is easy to see the similarity of SQP architectures, as described above, to 2.5D-IC, where we can view the C-chip as the interposer and the Q-chip as an active classical chip. Or, if we put some qubits on the C-chip, the architecture more closely resembles 3D-IC. Therefore, people may think that we can directly adopt the advanced packaging technologies used in the semiconductor industry during the scaling up of the SQP. However, since the working principles of a superconducting qubit and a transistor are not the same, there will be differences between the implementations of 3D integration technologies on these two kinds of chips.

The first difference is the usage of underfill [38]. The step after the flip-chip bonding for 2.5D-IC and 3D-IC generally includes injecting underfill between the chip and the interposer (or another chip) to strengthen the package's mechanical robustness and isolate the bumps from the environment. However, for SQPs, to fulfill the requirement that qubits should face air to avoid additional losses has blocked the potential usage of underfill. As a consequence, the mechanical robustness of the flip-chip integrated SQP may be lower than the classical semiconductor chips.

The second difference is the role of bumps and vias. For a transistor to work, it needs a stable current supply to be able to go through its source and drain. But for a superconducting qubit, it ideally only needs pulsed microwave signals through capacitively-coupled control lines as an energy supply to change its state. (A different case is frequency-tunable qubits [7] and couplers; they need constant current to generate fixed-amount magnetic flux to maintain their frequencies, but the current doesn't need to go through the qubits/couplers directly.) Therefore, the bumps in an SQP are not necessary to facilitate the delivery of energy. They can work only as mechanical support and galvanic connections between the Q- and C-chip. On the other hand, in addition to delivering microwave signals to the qubits, the vias also connect the chip's two sides galvanically and suppress the low-frequency chip mode that may act as the mediator of qubit crosstalk.

The third difference is in the property of components. From the eyes of an electrical engineer, there is no active component inside an SQP. A superconducting qubit/coupler is no more than a parallel combination of a capacitor and an inductor (with some non-linearity), which can still be categorized as passive elements [48]. Therefore, there is

no strict rule to forbid us from moving qubits, resonators, etc. between the Q-chip and the C-chip with adjusted designs. From this aspect, a flip-chip integrated SQP has an architecture that sits in the middle between 2.5D and 3D-IC architectures.

2.2.4 Implementation requirements

When we make an effort to introduce a new architecture to our SQPs, we always want such a change to improve some aspects of the SQP while not causing a severe hit on its other properties. 3D integration technologies are known to improve the scalability of an SQP, but to clear their way to becoming well-accepted technologies used, there are two obstacles that we have to overcome.

The first one is the maintenance of the SQP's performance. Since a qubit's surrounding electromagnetic environment will be altered under the new architecture, and will go through additional fabrication processes, it is then reasonable to ask whether the coherence of the qubits in the new SQP can be at least the same as that of the old, single-chip SQPs. The same question can also be asked for single-qubit and two-qubit gate fidelities. If the resulting performance of the SQP with the new architecture is worse than the old one, it means that either the introduced additional loss on the qubit is significant, or there are some unknown effects. Both cases will create more uncertainties regarding the relationship between the performance and the complexity level of the design, and in the face of large uncertainties, people usually hesitate to move forward.

The second obstacle is the reduced yield of the fabrication. In general, there is a cumulative decay of the yield of the final device if each fabrication step has a yield less than 100% [49]. To obtain an acceptable yield of the SQP in the end, adding one more fabrication step means that the yield of each previous step needs to improve accordingly. Since moving SQPs to the 3D architecture will certainly add multiple steps to the whole fabrication process, maintaining the original yield, or even trying to improve it, at the same time will be a very hard task, but it is necessary for keeping the overall fabrication cost down.

2.3 Discussion of the future trend

Similar to the development of semiconductor chips, the required density of bumps and vias on SQPs will increase dramatically in the future. Current superconducting qubits typically have a size at the order of hundreds of micrometers [50–52]. Taking its size as 500 μm, with 1000 qubits we will need at least 256 mm^2 of the chip area, which is already comparable to the size of modern processors, but later has billions of transistors [53]. If such a qubit size is kept constant, the chip area will be too large for hosting millions of qubits [24]. Therefore, reducing the qubit size is necessary, along with the pitch and the size of bumps and vias if 3D integration technologies are implemented. We can expect that the most advanced packaging technologies in 2.5D and 3D-IC will be adapted to SQPs to meet such needs accordingly.

On the other hand, as the number of qubits increases, so does the number of control lines for qubits and couplers within the SQP. However, when the SQP has millions of qubits, it is economically unrealistic to connect all these lines to the room temperature

microwave sources and measurement tools as we do in our current setups [32]. Efforts are already under way to reduce the number of I/O lines between ambient and cryogenic temperature by replacing room-temperature arbitrary waveform generators with digital logic units at cryogenic temperature close to the SQP [54, 55]. Following the trend, such a logic unit can be integrated together with the SQP within one package, similar to the system-in-a-package (SiP) technology in the semiconductor industry.

CHAPTER III

Flip-chip integrated SQPs

As a widely used packaging method in the semiconductor industry [36], flip-chip bonding of one chip onto another is a mature technology. However, unlike transistors in classical semiconductor processors, qubits inside the superconducting quantum processor (SQP) are much more sensitive to defects surrounding them, hence the performance of SQPs may suffer from the additional fabrication processes for flip-chip integration. In addition, one may ask whether a qubit's coherence can still be maintained when there is an another chip hovering a few micrometers above the qubit.

This chapter shows the performance of the essential components of a flip-chip integrated SQP. It will show that the performance of these components is similar to that of the single-chip architecture, demonstrating that the change in the electromagnetic (EM) environment and additional fabrication processes do not necessarily degrade the performance of an SQP, while transitioning to the flip-chip architecture can significantly improve its scalability.

3.1 3D structures in flip-chip integrated SQPs

Our flip-chip integrated SQPs consist of two chips. The top chip is called the qubit chip (Q-chip), which contains qubits and couplers that are the basis of SQPs. The bottom chip is called the control chip (C-chip), containing all the components and feedlines to control and read out the qubits. By separating these components into two chips, SQP's scalability can be improved by simplifying the signal routing and suppressing crosstalks among qubits and couplers. In our flip-chip architecture, the components on the surface of each chip face each other after flip-chip bonding the Q-chip onto the top of the C-chip. See Figure 3.1, which shows the cross-sectional schematic of a flip-chip integrated SQP.

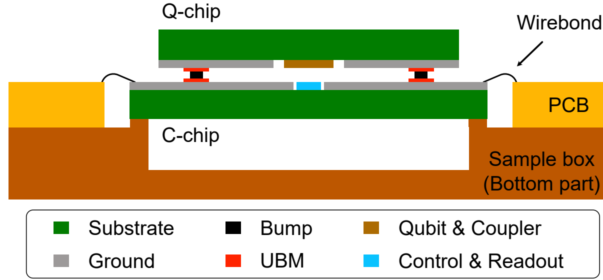


Figure 3.1: Cross-section of a flip-chip integrated SQP (not to scaled). Our flip-chip integrated SQP is connected to the outside world through a PCB via wirebonds and is encased in a sample box.

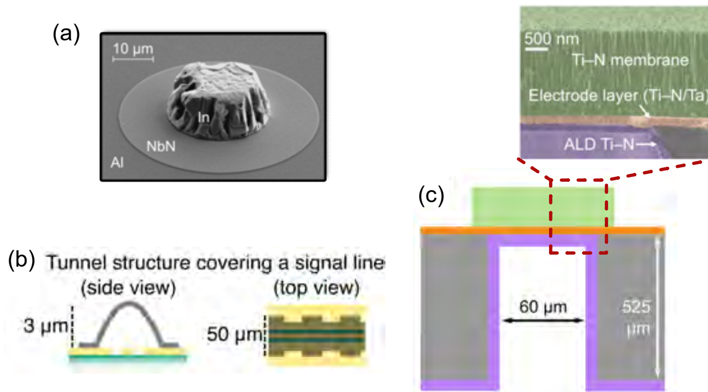


Figure 3.2: 3D structures for flip-chip integrated SQPs. (a) The scanning electron microscope (SEM) image of a bump sitting on top of UBM and the ground metal plane on the surface of the C-chip or Q-chip before the flip-chip bonding process. (b) Air-tunnel that covers the feedlines on C-chip. (c) a TSV that goes through the chip's substrate, with a zoom-in SEM image of each layer's material, fabricated by VTT (Paper B).

Figure 3.2 shows additional 3D structures that are implemented in our flip-chip integrated SQPs. Most of these structures are not new, as they have been widely used in semiconductor industries. The challenge lies in the choice of materials and fabrication processes that can maintain the performance of the SQP that has been achieved in the single-ship architecture [22].

Following is a brief description of 3D structures in flip-chip integrated SQPs:

- **Air-tunnel**: The air-tunnel is a 3D structure that is similar to airbridges used in single-chip SQPs [42]. Its usage is to suppress the feedlines' slotline mode and to stitch the separated ground plane of the chip to minimize the parasitic modes due to feedlines or other components on the chip, and shorten the return current path to minimize the flux crosstalk.

- **Bump**: Bumps act as the main supporting structure in the flip-chip integrated SQP to separate the top and the bottom chips. The interchip spacing is the most important parameter introduced by the flip-chip architecture since it can influence all components within the SQP. The height of a bump after bonding determines the interchip spacing. The bumps are generally chosen to be superconducting at cryogenic temperature, so they are also used as a conducting path between the ground planes of the two chips. A well-designed bump pattern across the SQP can reduce or even eliminate the need for air-tunnels in flip-chip integrated SQPs, as in Paper E. Bumps can also be used for signal routing from one chip to another chip [56].

- **Under-bump metalization (UBM)**: UBMs consist of one or several layers of thin films between the bump and the wiring layer. Their main purpose is as a diffusion barrier to avoid the direct contact between the bump material and the top metal layer of the chip, which may otherwise form intermetallic compounds [44, 56]. In addition, UBM can help with bump formation and ensure a galvanic connection between bumps and the chip's ground plane.

- **Through-silicon via (TSV)**: TSV is a vertical structure etched through the silicon substrate of the chip. By depositing a thin metal layer within the via, it can provide galvanic connectivity between the two sides of the substrate, such that signals can be routed between the two. By connecting both sides' ground planes, the fundamental frequency of the device's chip mode can be shifted well above the working frequencies of the SQP [56, 57]. It also can directly work as part of a qubit or a readout resonator of the SQP [58].

3.2 Fabrication process

In this section, we will briefly describe how flip-chip integrated SQPs have been made in our laboratory. A more detailed description of the fabrication process can be found in Paper A.

We use high-resistivity intrinsic silicon wafers as the substrate for both the Q-chip and the C-chip. We do the wafer cleaning process by going through Standard Cleaning 1, a generally used cleaning procedure for removing organic residue from silicon wafers, and HF dip for removing native oxide on the surface of the silicon wafers [59], and then immediately load the wafer into the evaporator. We deposit an aluminum (Al) thin film on top of the substrate, then sputter a niobium nitride (NbN) thin film on a patterned lift-off resist to form UBMs. The wiring layer of the chip is defined by etching out of the Al film.

We fabricate Josephson junctions using the patch-integrated cross-type technique [60] on the wafer containing the Q-chips. Next, indium (In) is evaporated on a patterned thick resist, and 7 µm-high In bumps are formed after resist lift-off. The Q-chip and C-chip are diced from the wafers, and the Q-chip is flip-chip bonded to the C-chip by mechanical compression by approximately 50% at room temperature.

The fabrication of TSVs was developed independently for a single-chip device and has not yet been integrated into the fabrication process of flip-chip integrated SQPs. The via holes of the TSVs are etched through the back side of the wafer and stopped by the thick titanium nitride (TiN) membranes patterned on the front side. The inner walls of the vias and the back side of the wafer are coated with TiN to enable conductivity between the front and the back side of the chip. A detailed fabrication process of TSVs can be found in Paper B.

3.3 Individual qubit performance

It is necessary to show that the additional fabrication processes of flip-chip integrated SQPs do not degrade qubit coherence compared to those of a single-chip architecture. The coherence of individual qubits in a flip-chip architecture, i.e. qubits not connected by couplers, is a good benchmark. Figure 3.3(a)-(b) shows the circuit diagram and the layout of such an individual qubit, together with its readout resonator.

Figure 3.3(c)-(b) shows qubits' relaxation time T_1 and decoherence time T_2^* that are comparable to what we usually achieve in the single-chip architecture [22].

To illustrate the reason why the flip-chip fabrication process has not degraded the qubit coherence, we conduct 2D EM simulations on the cross-section of a CPW line in both single-chip and flip-chip architecture. Since the majority of coherence loss of the qubit at cryogenic temperature comes from the two-level systems (TLS) [41] within the interfaces between metal thin film, substrate, and the vacuum (air), we can obtain a Q-factor of the cross-section by simulating the energy participation ratio of each interface and calculate it using the following equation [26]:

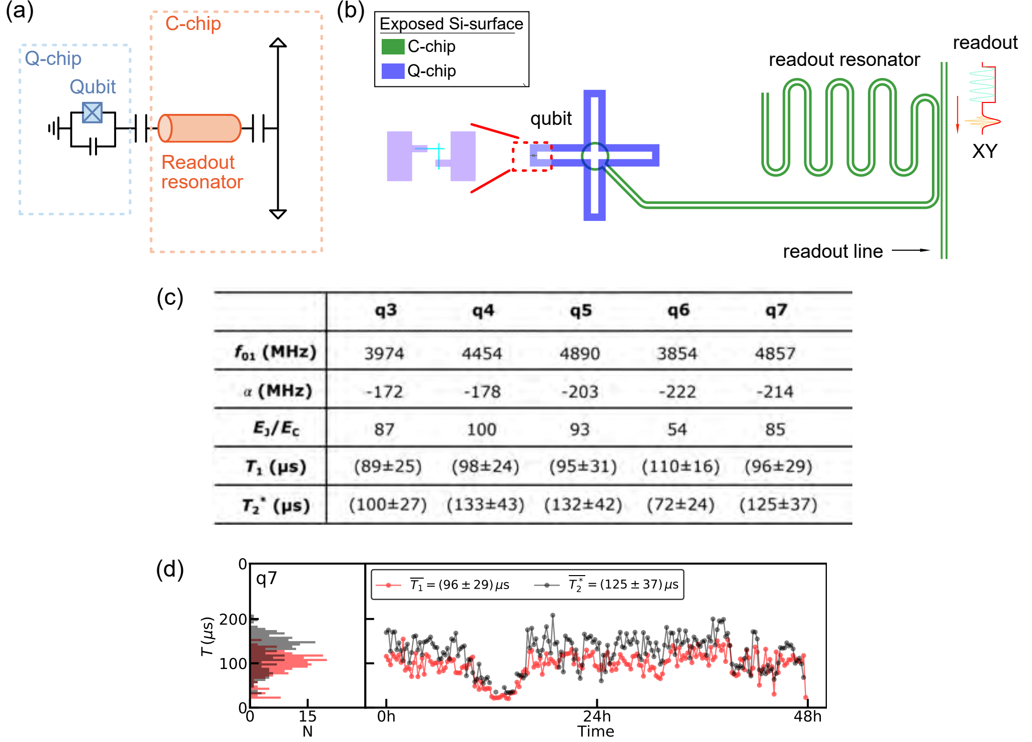
$$1/Q = \sum_i p_i \tan \delta_i, \tag{3.1}$$

where p_i is the simulated participation ratio for each domain i and $\tan \delta_i$ is the domain's loss tangent. The participation ratio p_i of a domain Ω_i is calculated by

$$p_i = \frac{w_i}{w}, \qquad w_i = \int_{\Omega_i} \epsilon_{\mathrm{r}}(\Omega_i) \left| \vec{E}(\vec{r}) \right|^2 d\vec{r}, \qquad w = \sum_i w_i, \tag{3.2}$$

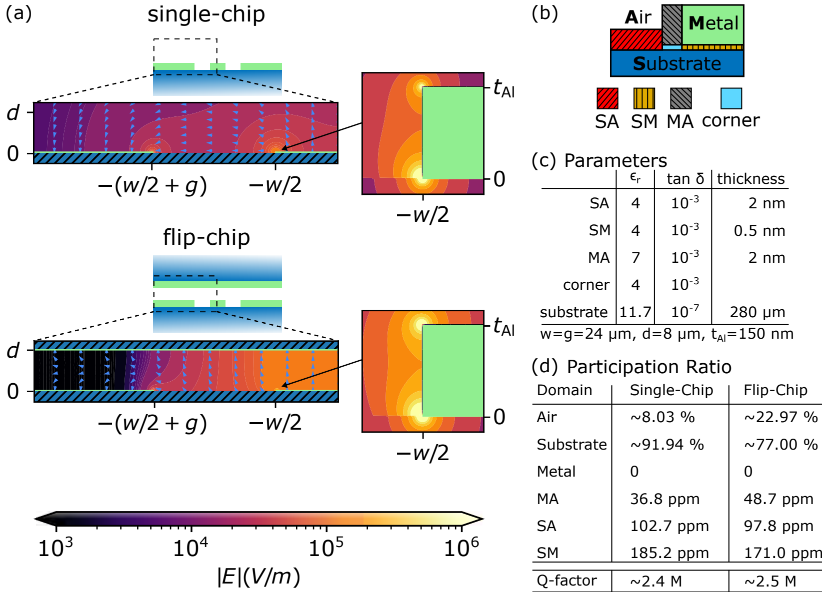
where $\epsilon_r(\Omega_i)$ is the relative permittivity of domain Ω_i, and $\vec{E}(\vec{r})$ is the simulated electric field (E-field).

Figure 3.4(a) shows the E-field distribution of both architectures. We found similar Q-factors of CPW lines in single- and flip-chip architectures, as shown in Figure 3.4(d), indicating that the total loss that a qubit experiences is similar within both architectures. The detailed simulation setup can be found in Paper A.



	q3	q4	q5	q6	q7
f_{01} (MHz)	3974	4454	4890	3854	4857
α (MHz)	-172	-178	-203	-222	-214
E_J/E_C	87	100	93	54	85
T_1 (μs)	(89±25)	(98±24)	(95±31)	(110±16)	(96±29)
T_2^* (μs)	(100±27)	(133±43)	(132±42)	(72±24)	(125±37)

Figure 3.3: (a) Circuit diagram of an individual qubit, separating Q-chip and C-chip. (b) The physical layout of the qubit, with a zoom-in illustration of the Josephson junction design. The qubit's control (XY) pulse and readout pulse are sent from its readout line. (c) Table of parameters of five measured individual qubits. f_{01} is qubit's frequency and α is the anharmonicity of the qubit. E_J is the Josephson energy and E_C is the charging energy [13]. (d) Example of qubit's coherence times T_1 and T_2^* fluctuations over time of one of the qubits, with accumulated histograms.



	ϵ_r	tan δ	thickness
SA	4	10^{-3}	2 nm
SM	4	10^{-3}	0.5 nm
MA	7	10^{-3}	2 nm
corner	4	10^{-3}	
substrate	11.7	10^{-7}	280 μm

Domain	Single-Chip	Flip-Chip
Air	~8.03 %	~22.97 %
Substrate	~91.94 %	~77.00 %
Metal	0	0
MA	36.8 ppm	48.7 ppm
SA	102.7 ppm	97.8 ppm
SM	185.2 ppm	171.0 ppm
Q-factor	~2.4 M	~2.5 M

Figure 3.4: 2D simulation of the E-field within the single-chip and flip-chip architecture. (a) Contour plots of the E-field magnitude for the same CPW geometry in both single-chip and flip-chip environments. The arrow indicates the direction of the E-field. (b) Designation of the various domains for the participation ratio simulation. (c) Parameters used in the simulation. (d) Calculated total Q-factor and participation ratios of the different domains for both single-chip and flip-chip architectures.

3.4 Two-qubit pair as a unit cell

An SQP contains not only qubits but also couplers between the two adjacent qubits, which enable two-qubit interaction. A proper design of a two-qubit pair containing two qubits and one coupler, together with their control and readout components on the C-chip, can work as a basic unit cell to construct SQPs with arbitrarily large numbers of interconnected qubits. Figure 3.5(a) and (b) show the circuit diagram and the physical layout for a two-qubit pair in flip-chip architecture.

As illustrated by Figure 3.5(b), for such a coupled two-qubit design, the single-qubit gate is performed via the drive line beneath the qubit. Since we use the parametric controlled-phase gate [61, 62] as our native two-qubit gate, a flux-tunable qubit is used as a coupler, and we control the coupler through the current loop at the end of the flux line. The qubit's states are read out through their corresponding readout resonators. Figure 3.5(c)-(d) show the fidelities of both single- and two-qubit gates using randomized benchmarking experiments [63], and their values are comparable to what we can achieve in the single-chip architecture [22].

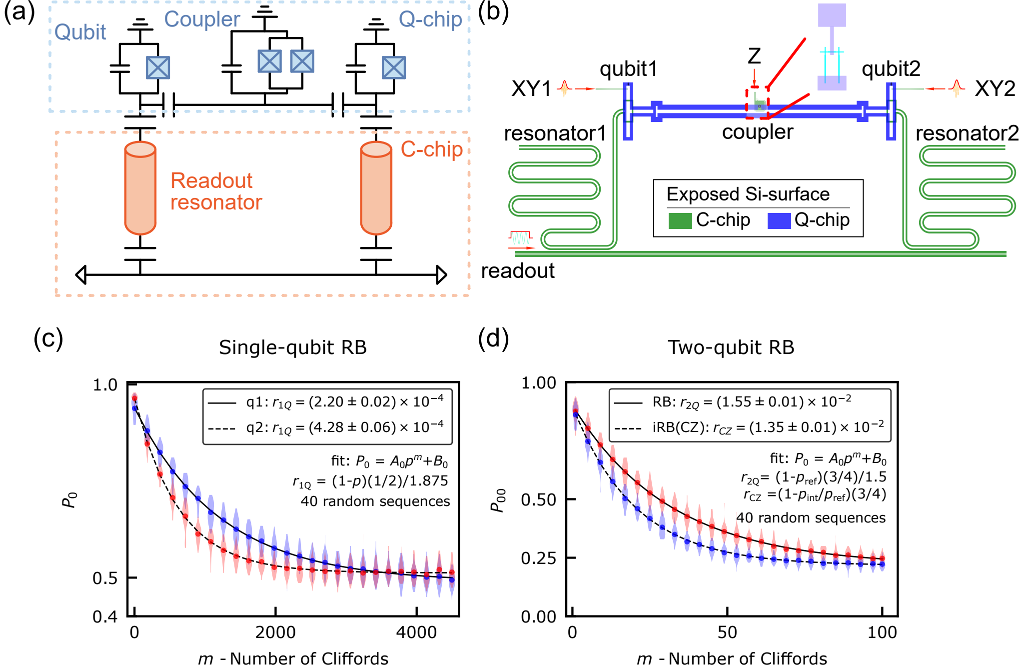


Figure 3.5: (a) Circuit diagram of the two-qubit pair design, separating Q-chip and C-chip. (b) The physical layout of a two-qubit pair, with a zoom-in illustration of the SQUID (superconducting quantum interference device that made of a loop of two parallel Josephson junctions such that the coupler's frequency is tunable by the magnetic flux.) of the coupler. The control (XY) pulse and readout pulse are sent to the qubit through the drive line and the readout line separately. The coupler is controlled through a flux (Z) line. (c) Measured single-qubit gate fidelity ($\approx 99.95\%$) and (d) two-qubit gate fidelity ($\approx 98.65\%$) using randomized benchmarking (RB) and interleaved randomized benchmarking (iRB).

3.5 Purcell filter for qubit readout

To read out the state of a qubit, the qubit must couple to a readout line. However, such coupling creates an energy leakage channel for the qubit. The stronger the coupling for faster measurement, the worse the leakage will be. Generally, to suppress such leakage, and at the same time still realize fast measurement of the qubit, a readout resonator is dispersively coupled to the qubit [13]. The measurement of the qubit state is then realized by monitoring the resonator's qubit-state-dependent resonant frequency through the readout line that is coupled to the resonator only. However, though suppressed, the energy leakage of the qubit cannot be completely eliminated. The decaying mechanism of qubit excitation through its readout resonator is called Purcell decay [64], which is in essence the excited qubit's spontaneous emission of a photon into the readout resonator.

To further suppress the qubit energy leakage, a second resonator is introduced between the first resonator and its readout line. We call the second resonator the Purcell filter [65].

When multiple qubits and their readout resonators exist, each qubit's readout resonator can couple to one Purcell filter that has the same resonant frequency and bandwidth as the readout resonator [66]. Alternatively, the Purcell filter can be a long single-pole resonator that simultaneously couples to multiple readout resonators, hence multiple qubits at the same time, with its central resonant frequency near the readout resonators' frequencies and a large bandwidth to include all the coupled resonators [67], while the qubit frequencies lie at the high attenuation part of the Purcell filter transmission spectrum. Such a Purcell filter can be called a common Purcell filter. Detailed theory describing how Purcell filter works can be found in Paper C.

Both types of Purcell filters can be implemented in flip-chip integrated SQPs. We choose to use the common Purcell filter because of its simplicity in layout and less sensitivity to the variations in the fabrication process. Figure 3.6 shows the circuit diagram and the layout design of such a Purcell filter, together with a parameter optimization diagram for one qubit coupled to this Purcell filter.

Benefiting from the further suppression of the qubit's Purcell decay, we can now design the qubit to be much more strongly coupled to its readout resonator, such that the readout speed can be significantly increased while at the same time retaining the qubit's high coherence. The transmission spectrum of the Purcell filter together with five strongly coupled readout resonators is shown in Figure 3.7. Table 3.1 shows the measurement time (τ_m) of the Purcell filter-protected qubits (Q16-Q20) when their two-state readout assignment fidelity reaches maximum ($\mathcal{F}_a^{max}$), compared to those qubits without Purcell filter (Q21-Q25). In general, qubits protected by the Purcell filter exhibit higher readout fidelity with shorter measurement time.

It is noticeable that both kinds of qubits have similar coherence times (T_1, T_2^*) that are much smaller than Purcell limit T_p, the relaxation time limit of the qubit due to Purcell decay, meaning their coherence times are limited by loss channels other than the Purcell decay. To illustrate, Figure 3.8 shows that the T_1 of the qubit is capped by losses from other channels if T_p is much higher than T_{other}, where T_{other} is qubit's relaxation time limit if it has no Purcell decay. It is worth emphasizing that the goal of Purcell filters in SQPs is to retain the qubit coherence while improving the qubit's readout fidelity and shortening the measurement time by increasing the coupling rate $g_{\text{q,r}}$ between the qubit and the readout resonator. To this end, our implementation of the Purcell filter in the flip-chip integrated SQP has fulfilled this goal.

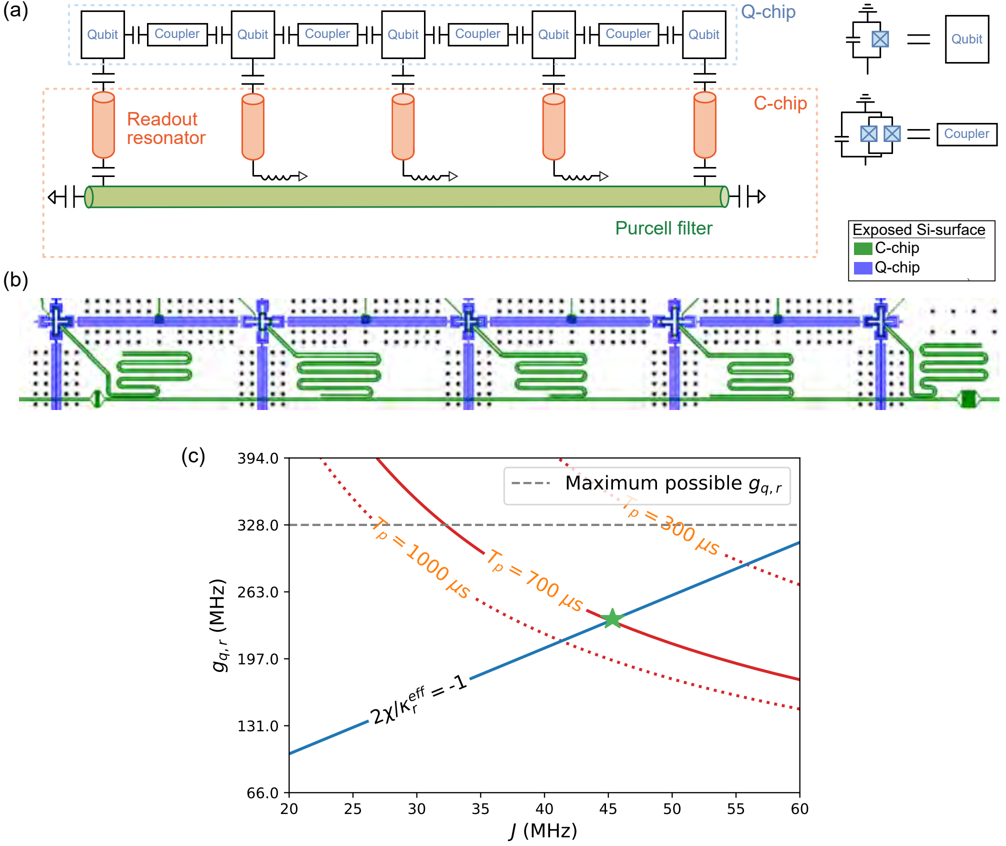


Figure 3.6: (a) Circuit diagram and (b) physical layout example of five qubits and readout resonators coupled to a common Purcell filter. The Purcell filter is a half-wave ($\lambda/2$) resonator defined by two capacitors which impose open boundary conditions at the input and output ports of the readout line. (c) Parameter optimization diagram of one of qubits coupled to the Purcell filter to determine the qubit-resonator coupling rate $g_{\mathrm{q,r}}$ and resonator-filter coupling rate J that satisfy the optimal readout condition $2\chi/\kappa_r^{\mathrm{eff}} = -1$ [13], and the resulting Purcell limit T_p. Here χ is the dispersive shift of the readout resonator, and κ_r^{eff} is the effective leakage rate of the readout resonator in the presence of the Purcell filter. The dashed line indicates the maximum possible $g_{\mathrm{q,r}}$ our layout design can achieve.

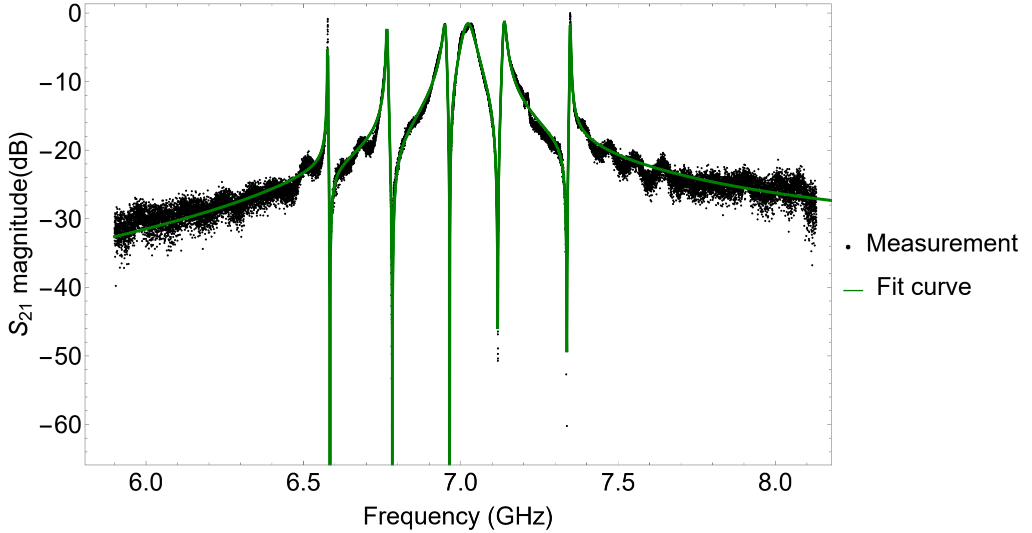


Figure 3.7: The transmission, S_{21}, spectrum of the readout line with a common Purcell filer, coupled to five readout resonators and qubits. The fit function of the curve is theoretically derived in Paper C. Note that the qubit frequencies are below 6 GHz.

Qubit Nr.	f_{01} (MHz)	$g_{\mathrm{q,r}}/2\pi$ (MHz)	T_p (µs)	T_1 (µs)	T_2^* (µs)	$\mathcal{F}_a^{max}$ (%)	τ_m (µs)
Q16	4208	126	1535	36	46	99.44	0.84
Q17	5079	109	1937	36	47	99.05	1.04
Q18	4398	278	1369	44	72	99.07	0.68
Q19	4656	220	2175	50	40	96.83	0.66
Q20	4033	255	681	35	41	99.34	0.78
Q21	4964	53	513	26	51	96.04	2.08
Q22	4420	55	617	57	65	98.19	1.75
Q23	4855	46	587	31	56	97.17	2.1
Q24	4167	66	506	47	60	98.58	1.97
Q25	4766	58	583	27	26	78.59	1.07

Table 3.1: Qubit characterization results of two readout lines, in which one line is replaced by a common Purcell filter shown in Figure 3.6(b) and qubits on this line (Q16-Q20) are strongly coupled to the readout resonator, while another line has no the Purcell filter and qubits (Q21-Q25) are weakly coupled.

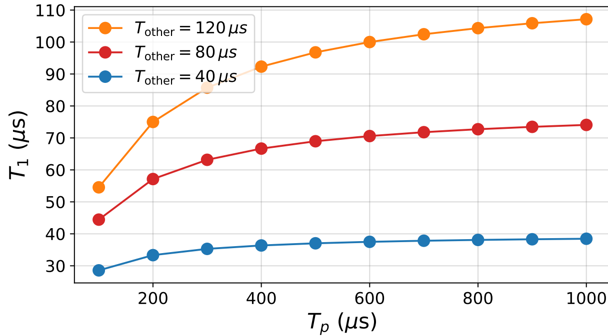


Figure 3.8: Expected qubit relaxation time T_1 at different values of the Purcell limit T_p, considering losses from other channels that also limit the qubit's relaxation time to T_{other}. The resulting qubit relaxation time is calculated by $1/T_1 = 1/T_p + 1/T_{\mathrm{other}}$ [26].

3.6 Interchip spacing variation effect

In the flip-chip architecture, one of the most important parameters is the interchip spacing between the C-chip and the Q-chip. Since most of the components have a relatively large footprint on the chip, the variation of the interchip spacing can significantly affect the performance of the SQP components through the change of their self and mutual capacitances and inductances. This section shows how the interchip spacing affects these components under global and local variations of the interchip spacing across the processor. Since in our current multi-qubit, flip-chip integrated SQP design, which will be discussed in Chapter 4, all components on one chip face the ground metal plane of another chip, we only discuss such a scenario in this section. Discussions on the scenario in which the readout resonators face the silicon substrate of the opposite chip will be discussed in Chapter 5.

3.6.1 Global variation

When there is a variation in the calibration of the flip-chip bonding force, or other factors resulting in a global variation of the height of the formed bumps prior to the bonding step, the fabricated flip-chip integrated SQP may have a homogeneous deviation of interchip spacing from the target value across the whole processor. This global shift will cause all components of the SQP to have certain deviations from their designed parameters.

A deviation of the interchip spacing changes a component's capacitances and geometric inductances to its surrounding ground plane and to other nearby components, which then change the component's other related design parameters, resulting in a deviation from its intended performance.

For CPW lines that make up the feedlines (qubit's drive line and readout line, coupler's flux line), readout resonators, and Purcell filters, the variation of interchip spacing propagates to the variation of the characteristic impedance of the CPW lines. Typically, the characteristic impedance of the feedlines is designed to be the same as lines outside

the SQP. The mismatch of the impedance at the boundary of the SQP, if the deviation of interchip spacing is too large, can cause significant input and output signal reflections [68], reducing the effectiveness of qubit control and readout. Figure 3.9 shows the change of the readout lines' characteristic impedance as the interchip spacing changes. For readout resonators and Purcell filters, a change in characteristic impedance results in a change in their resonant frequencies and, thus, their coupling rates to other components. A detailed derivation of the effect of interchip spacing on CPW lines and thus resonators will be discussed in Chapter 5.

The interchip spacing that a qubit sees affects the qubit's frequency f_{01} and its coupling strength to its readout resonator $g_{\mathrm{q,r}}$ and drive line, as the capacitance between the qubit island and the C-chip's metal film on the surface is directly correlated with the interchip spacing. Figure 3.10 shows the effect of the interchip spacing deviation on qubits' parameters, in particular to its Purcell limit time (T_P) that sets the upper limit to the qubit's lifetime T_1 in the absence of other loss channels.

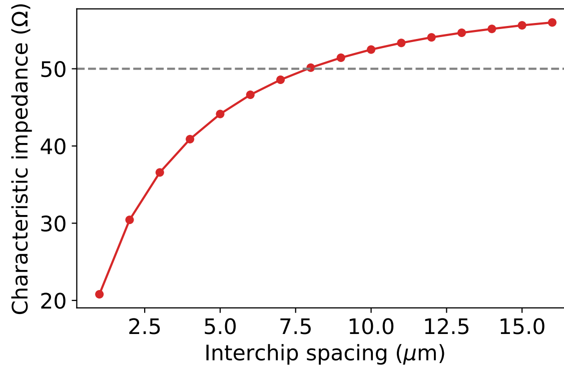


Figure 3.9: Feedlines realized by CPW lines have characteristic impedance changing with the interchip spacing in the flip-chip architecture. The CPW line here has a center width of 9 µm and a gap of 10 µm. The gray dashed line indicates the characteristic impedance when the interchip spacing is at its target value of 8 µm.

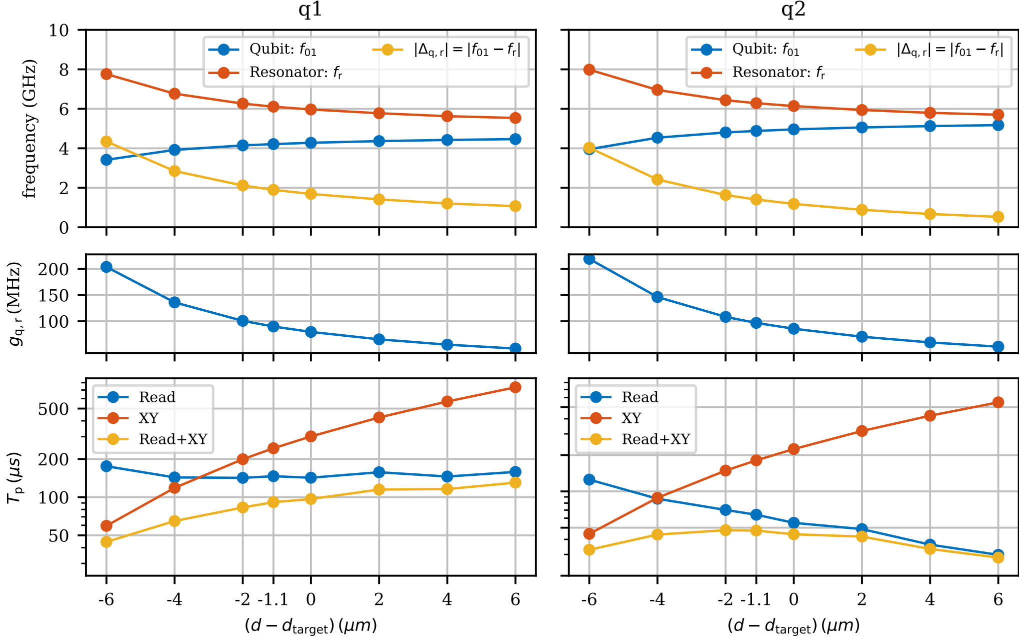


Figure 3.10: Qubit parameters change with the interchip spacing (d) when it deviates from the target value (d_{target}). The two qubits are from Figure 3.5(b) and the low Purcell limit time (T_P) of qubit 2 at d_{target} is due to higher-than-expected qubit frequency (f_{01}) resulted from deviated Josephson junction size. f_r is the resonant frequency of readout resonator. Qubit's T_1 limit due to decay through the drive (XY) line is also calculated and combined with qubit's Purcell decay within the figure.

3.6.2 Local variation

The non-uniformity of the bump height across the chip (chip tilt) and the chip's own non-flatness (chip warp) can cause local variation over the flip-chip integrated SQP, meaning that depending on each component's location, a different interchip spacing is seen by the component, causing a mismatch of design parameters between two adjacent components. In addition, for components with large footprints, such as the common Purcell filter, the prediction of its parameters becomes much harder.

For example, under certain conditions, the chip tilt is detrimental to the multiplexing readout of qubits. As the number of qubits increases with the scale of the SQP, more and more qubits must be read out through their readout resonators that are coupled to a single readout line. Due to the limited bandwidth of arbitrary waveform generators, the resonant frequencies of these readout resonators need to be arranged within a small bandwidth. Currently, a 200 MHz frequency separation is used between different resonators. A 1 µm change of the interchip spacing roughly corresponds to a 100 MHz shift of the resonator frequency (details will be shown in Chapter 5 and can be found in Paper D). This means that if the chip tilt causes two resonators that are close in frequency to have a 2 µm interchip spacing difference, a frequency collision is expected to happen.

3.7 Further developments in flip-chip architecture

The full integration of TSVs is expected to be the next step in the development of the flip-chip architecture to improve ground plane stitching for the suppression of chip mode and parasitic modes. In addition, TSVs can be used for signal routing with the utilization of a redistribution layer (RDL) as a third layer below the C-chip [56] to simplify the wiring layer on the C-chip. The connection between TSVs and RDL can be realized by an additional sets of bumps between the C-chip and RDL.

Another promising development is to flip-chip bond multiple Q-chips (chiplets) to a single C-chip in a multi-chip module (MCM) [69], similar to 2.5D-IC architecture in the semiconductor industry. The major advantage of such a packaging method is to reduce the variation of Josephson junction resistance and the number of defects within the SQP. As the number of qubits an SQP is expected to host keeps increasing, so does the size of the Q-chip. Unfortunately, fabrication variation and the number of defects also increase with the size of the Q-chip. This results in a lower yield per fabrication round and translates to a too-high cost for making large SQPs. Treating each Q-chip as a chiplet with a small number of qubits can then enable us to scale up the SQP while keeping or even shrinking the size of the Q-chip. Supported by pre-bonding verification steps (Josephson junction resistance probe and defect inspection), we can discard those Q-chips with defects or those that are out of the parameter tolerance range. In this way, building a large SQP by flip-chip bonding of multiple 'good' small Q-chips on a large C-chip is feasible.

CHAPTER IV

Design of a multi-qubit, flip-chip integrated SQP

Similarly to developing a classical semiconductor processor, the full design cycle of an SQP goes through four stages: **chip specification**, **design**, **fabrication**, and **characterization**. At the stage of chip specification, before any design effort is devoted to our SQP, we must determine how many qubits and couplers should be placed inside the SQP. This needs to consider the SQP's intended usage, such as target algorithms and applications, as well as general constraints like the chip size, available I/O ports, etc. As an example, our largest flip-chip integrated SQPs to date have 25 qubits and 40 couplers, based on the footprint of the two-qubit pair shown in Chapter 3 and our conventional Q-chip size of 12 mm $\times$ 12 mm. The number of required I/O ports for control and readout lines is 75, which is within the number of ports (80) our current PCB and sample holder can provide.

After chip specification, design can start. This chapter describes the design process of a scaled-up multi-qubit flip-chip integrated SQP and, in the end, discusses necessary improvements to the current design process if we aim to scale up the SQP further.

4.1 General design workflow

The general design workflow of a multi-qubit SQP is shown in Figure 4.1.

In brief, the chip design of SQP can be divided into five steps: **architecture design**, which determines how qubits are interconnected via couplers to form a network; **parameter design**, which, at the circuit level, specifies the target parameters of each component to have optimal performance for intended usages; **layout design**, which determines the physical layout and geometrical placement of components that meet the target parameters and the feedline routing; **design verification**, which includes design review and design rule checking, to spot any design error that may cause deviated parameters and fabrication in-compatibility; and in the end, **tapeout**, which is the submission of the final layout of the SQP to fabrication.

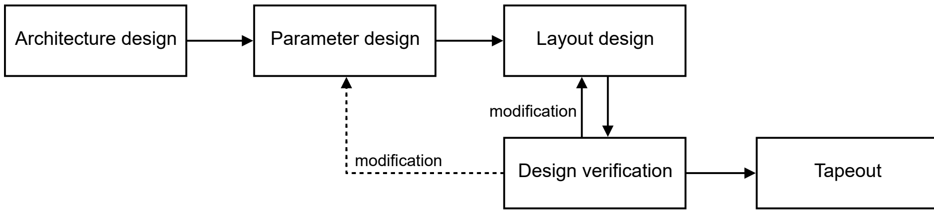


Figure 4.1: General design workflow of a superconducting quantum processor. Note that design verification may result in modification of layout design and even parameter design.

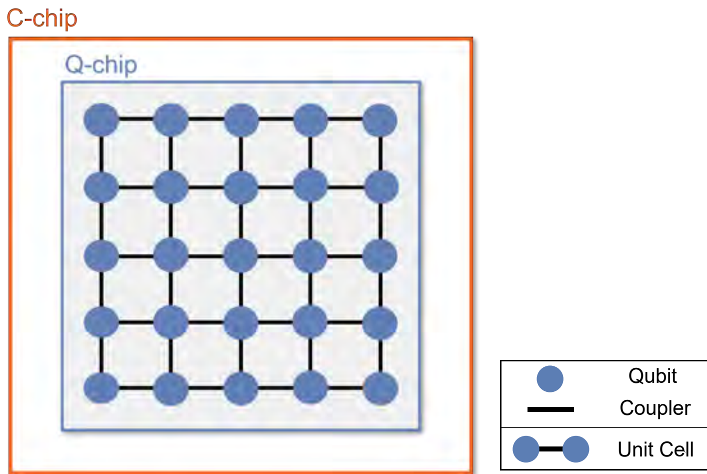


Figure 4.2: Architecture design chosen for 25-qubit flip-chip integrated SQP at out lab.

4.2 Architecture design

After specifying the number of qubits and couplers, we need to determine their relative positions within the SQP, which we call architecture design. To realize fault-tolerant quantum computing [6], the architecture of an SQP is typically set by the quantum error-correcting code we want to implement on top of it. For example, qubits placed in a square grid with each qubit having four nearest neighbors are generally used for realizing the surface code [2, 43, 70, 71].

In our flip-chip integrated SQPs, we choose to use the square grid architecture, as it can be flexible enough to implement both error-correction codes and quantum simulations [72]. We will have qubits connected with the four nearest neighbors except those on the edges, as shown in Figure 4.2.

Component	Parameters
Qubit	- Frequency - Anharmonicity - Purcell limit due to readout resonator and drive line - Expected coherence times due to various loss channels
Coupler	- Idle frequency at zero bias - Frequency tunable range by flux line - Anharmonicity - Qubit-coupler coupling rate
Readout resonator	- Resonant frequency - Dispersive shift due to qubit - Leakage rate to readout line - Critical photon number
Purcell filter	- Resonant frequency - Leakage rate to readout line - Filter-readout resonator coupling rate

Table 4.1: Component parameters of our flip-chip integrated SQPs for optimization at the parameter design step. Note that not all of the parameters are independent, and other parameters not shown here can be derived from these parameters.

4.3 Parameter design

The parameters that need to be designed at this step are those that directly constitute the SQP's circuit Hamiltonian and those that affect measurements. In Table 4.1, we list the major parameters of various components as target design parameters.

In general, the parameter design step of the SQP can be divided into three parts. We first consider the parameter constraints at the unit cell level (the two-qubit pair in our case) to establish the possible parameter range of each unit cell. The second part is to consider the placement of the different cells' parameters, with constraints ensuring that interference between each cell is maximally suppressed. This usually results in a narrower possible parameter range. Finally, we optimize the parameters of each cell within the parameter range to achieve the best performance at both the cell and chip levels.

The following subsections illustrate these three parts for our 25-qubit SQPs.

4.3.1 Constraints on the unit cell level

The exact parameter constraints of a unit cell, like the two-qubit pair for our SQPs, depend on the type of qubits within it and the chosen native quantum gates we want to implement. For our flip-chip integrated SQPs, we use fixed-frequency transmon qubits and flux-tunable couplers to realize the parametric controlled-phase gate [61, 62], which sets the following parameter constraints:

- Qubit's frequency should be around 4-5 GHz.
- Qubit's anharmonicity should be around 200 MHz.
- Readout resonator's resonant frequency should be around 6-7 GHz, so that qubit-readout resonator frequency detuning is around 2 GHz.
- Ratio between the readout resonator's dispersive shift due to qubit and leakage rate to readout line should be 0.5.
- Qubit-coupler coupling rate should be around 30 MHz.
- Coupler's idle frequency at zero bias should be above the highest readout resonator's resonant frequency when doing multiplexing readout.
- Coupler's frequency tunable range from idle frequency should be able to reach qubit's frequency.

4.3.2 Constraints on the chip level

There are two types of chip-level constraints: one concerns interference between different qubits, while another between readout resonators during the multiplexing readout.

Frequency collisions are detrimental to the performance of SQPs as they introduce significant crosstalk between the two qubits and coherent errors when implementing two-qubit gates [73]. To avoid frequency collisions between qubits and during two-qubit gate implementation, the separation between the qubit frequencies should be larger than the fabrication variation.

In the presence of a common Purcell filter, as shown in Chapter 3, readout resonators' resonant frequencies need to be as close as possible to the central resonant frequency of the Purcell filter. This is for obtaining a large effective leakage rate of the readout resonator to the readout line. However, to avoid frequency collisions (bandwidth overlapping) between different readout resonators, the readout resonator's frequency separation is constrained by the fabrication variation. For flip-chip integrated SQPs, a frequency variation of ± 100 MHz is caused by the chip spacing variation of 1 μm as will be shown in Chapter 5. In addition, the frequency variation due to over- and under-etching of CPW lines (central lines and ground edges) is within ± 10 MHz.

4.3.3 Parameter optimization

The aim of parameter optimization is to maximize the accuracy of qubit control and measurement so that the intended quantum algorithms can be successfully implemented. For our qubit architecture, this means maximizing the fidelity of single-qubit gate, two-qubit gate, and qubit's readout fidelity.

The optimization of quantum gates is itself an active research area. Techniques to implement control pulses to qubits and couplers keep advancing [74–76], along with them are different optimal sets of the qubit/coupler parameters for these techniques [77]. However, as the first requirement, the coherence of qubits should be as high as possible. The coherence of a qubit directly affects the fidelity of the single-qubit and two-qubit gates

due to the accumulated errors during the gate operation time [78]. That means that while optimizing various parameters, the influence of these parameters on qubit coherence must be taken into account. Another major optimization objective is to suppress both XY and ZZ crosstalk between qubits. The XY crosstalk is determined by the direct capacitance between two adjacent qubits and their frequencies as discussed in Paper E. To reduce the XY crosstalk, we can either decrease the direct capacitance between two qubits by moving them further away during layout design or increase the frequency detuning between two qubits. The ZZ crosstalk is the result of the interaction between qubits in which the phase accumulation of one qubit is conditional on the state of another qubit. There is an optimized parameter region of qubits and couplers such that the ZZ crosstalk can be minimized and at the same time fast and high-fidelity two-qubit gates are achievable [79].

The readout fidelity of a qubit is affected by its coherence, readout pulse duration, and accumulated signal-to-noise ratio (SNR) at the end of the pulse. An optimal set of parameters should not only retain qubit coherence, but also maximize the accumulation rate of the SNR [80]. The purpose of introducing the Purcell filter is, in the end, to further increase the accumulation rate of SNR while not degrading qubit coherence.

An example design parameter set of our 25-qubit SQPs can be found in Reference [81].

4.4 Layout design

After determining the target design parameters, we can start drawing the SQP's physical layout. For the flip-chip integrated SQPs, we first design the layout of the Q-chip, which contains only qubits and couplers. Then we determine the bump pattern all over the chip. Next, we settle the routing of control and readout feedlines on the C-chip from the launch pads. Finally, we decide on the layout of readout resonators and optionally Purcell filters.

If the unit cell of the processor has been demonstrated to be workable, as the two-qubit pair shown in Chapter 3, the exact layout of different unit cells that satisfy the target parameters can be determined via analytical calculation, EM simulation, and previous successful testing fabrications. We do 3D EM simulations to determine each two-qubit pair's layout at different physical locations that satisfy the target capacitances between the qubit and the coupler, the qubit and its readout resonator, and the qubit/coupler's self-capacitances. The area of the Josephson junction is determined by a calibration curve (junction lead width vs. resistance) obtained from the room-temperature normal resistance measurement of thousands of junctions fabricated on a wafer [73]. During the EM simulation, we also determine the end position of the qubit drive line to achieve target capacitance to the qubit and the flux loop size to achieve target mutual inductance to the coupler's SQUID. The resulting larger unit cell of our flip-chip integrated SQPs is shown in Figure 4.3.

The bump and UBM pattern across the processor is determined with three constraints. First, the number of bumps has an upper limit because the flip-chip bonding needs to compress these bumps, but the force of the flip-chip bonder has a maximum achievable value (our available flip-chip bonder can compress 2000 bumps). The second constraint is that the separation of two bumps should be larger than the width of the feedlines (including air-tunnels), and the placement of bumps should result in a large separation

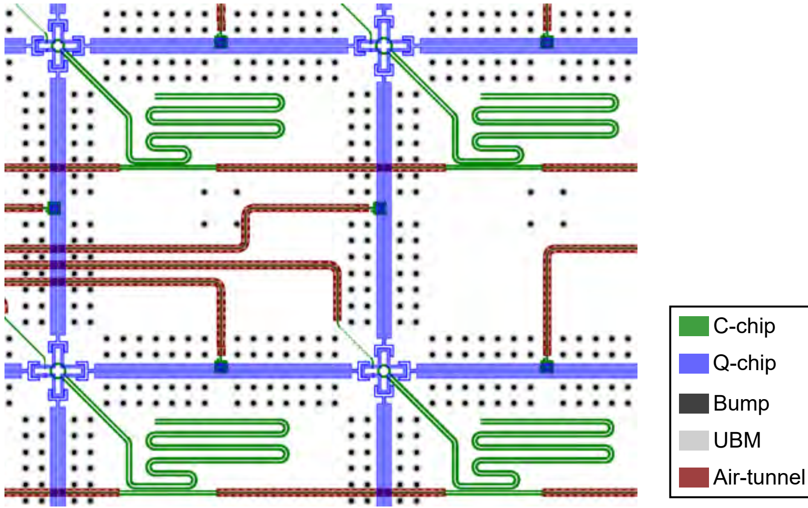


Figure 4.3: Part of our multi-qubit, flip-chip integrated SQP layout with 4 two-qubit pairs connected to each other.

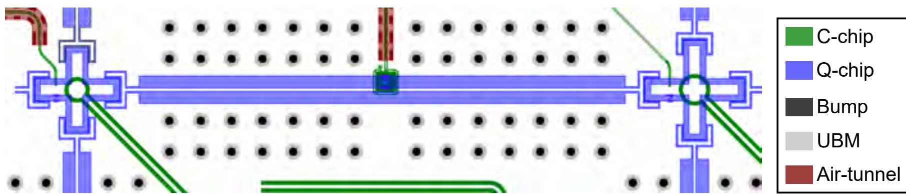


Figure 4.4: Layout of one of the two-qubit pairs, in which bumps are placed around the coupler to stitch the Q-chip ground plane divided by the coupler.

between feedlines, such that the crosstalk between adjacent feedlines can be suppressed (Paper E). The third constraint is to make sure the bump pattern satisfies the symmetry of a rectangle, such that the chip tilt due to uneven counter-compression force from the bumps during the flip-chip bonding can be minimized. In addition, the bumps should be placed around the couplers so that the ground plane sections of the Q-chip can be stitched together through the ground plane of the C-chip. Figure 4.4 shows how bumps are placed around a coupler.

Control lines are routed from their endpoints, i.e., from the drive line's open end and the flux loop. We set as a principle that the drive lines and flux lines should be alternately juxtaposed in parallel. This is because the signal frequencies through the drive lines (4-5 GHz) are different from signals through flux lines (around 200 MHz + DC current). By avoiding two drive lines or flux lines being routed side by side, crosstalk between the lines can be suppressed. On the other hand, readout lines are routed within the C-chip area first which is covered by the Q-chip, at positions close to the qubits, to leave enough routing space for the control lines of the qubits and couplers in the next row. The routing of feedlines from the Q-chip area's edge to the C-chip's launch pads is

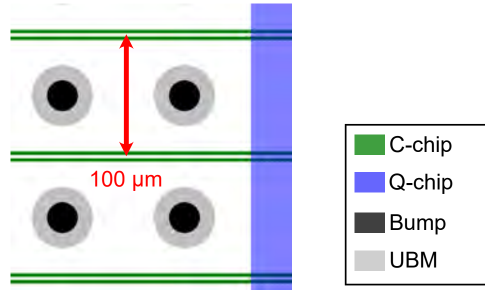


Figure 4.5: Layout of feedlines before crossing the coupler, with the air-tunnels layer hidden for a clear view. The separation between two nearby feedlines is 100 µm.

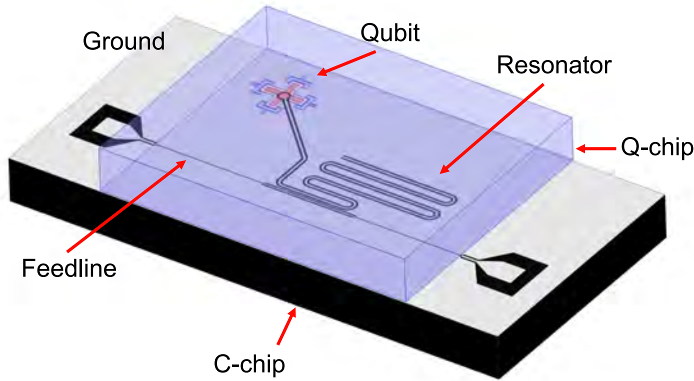


Figure 4.6: Readout resonator's 3D model in flip-chip architecture. The model is created in an EM simulation software [82].

flexible and depends on the available area of the C-chip outside the Q-chip's covered area. The overall constraint of the routing is that the separation of two nearby feedlines should be larger than 100 µm to minimize the crosstalk, as shown in Figure 4.5

The layout design of the readout resonators is done separately from that of the qubits since the resonator's target parameters are obtained through the simulated scattering matrix at GHz frequencies, while qubit capacitances are obtained by electrostatic simulations. We use analytical calculation and EM simulations to obtain the readout resonator's resonant frequency and coupling quality factor to its readout line. Figure 4.6 shows the simulation model of one of the readout resonators. We repeat such simulation for each different layout design of the qubit-readout resonator pair to ensure meeting the target design parameters. To include Purcell filters, we replace part of the readout line with a $\lambda/2$ resonator and simulate its scattering matrix accordingly, as shown in Chapter 3.

We parameterize the layout design of qubits, couplers, and readout resonators, not only within the 3D simulation models, but also within the component modules of Qiskit Metal [83]. A 25-qubit layout is then automatically generated by specifying layout

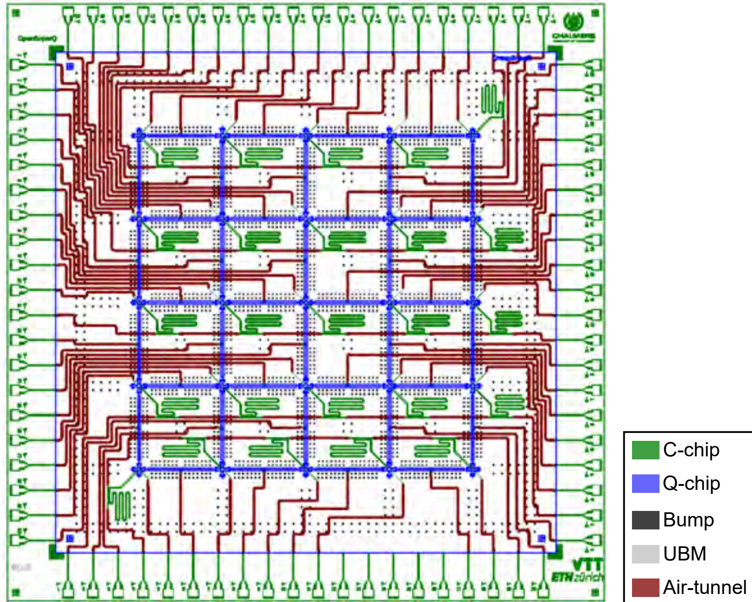


Figure 4.7: The design layout of one of our 25-qubit flip-chip integrated SQPs ready for fabrication.

parameter sets in a configuration file. To route feedlines within the layout, we define anchor points of each feedline, where each anchor point indicates that the feedline should have a 90-degree turn. All feedlines are then generated by Qiskit Metal by following these anchor points. The turning radius of the feedline is set to be 10 times of the CPW's center conductor width to avoid impedance mismatch and radiation at sharp bends [68]. If the air-tunnels following the feedlines are needed, their layout designs are done in L-Edit [84] manually. Flux holes, bump and UBM patterns on ground planes of the C- and Q-chip are also drawn within L-Edit. More details of the layout drawing workflow can be seen in the following subsection.

Figure 4.7 shows the C-chip and Q-chip layout of a 25-qubit SQP ready for fabrication. Except for the modified readout resonator position for the top-right and bottom-row qubits due to routing space constraints, the layout of this SQP is very modular, so a basic design cell can easily be identified. What's more, the routing of the feedlines within the Q-chip area is repeatable for center rows except for the top and bottom rows. This illustrates the scalability and modularity of our flip-chip integrated SQP layout design.

4.4.1 Layout drawing workflow of an SQP

In this section, we will briefly describe the layout drawing workflow of an SQP in our lab. After determining the components' geometry through 3D EM simulations, the next step is to draw out these components to constitute a completed SQP layout design ready for its fabrication.

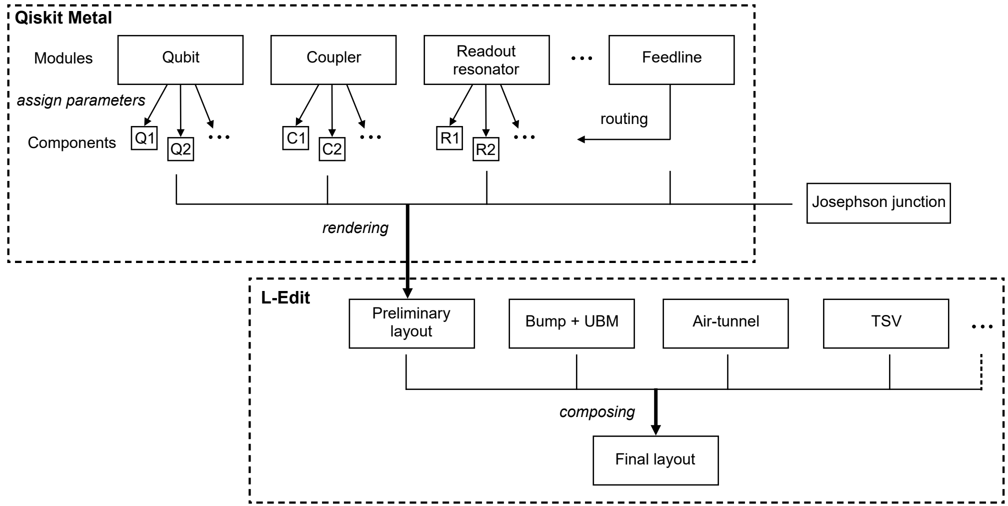


Figure 4.8: Layout design workflow of a multi-qubit SQP in our lab.

We use Qiskit Metal [83] as the main software for our layout drawing. We first need to write component modules describing the geometry of each component, such as qubits, couplers, readout resonators, etc. Then we write a function to assign determined geometric parameters to each component based on their physical locations on the SQP, and we write another module containing manually assigned anchor points to describe the routing path of feedlines from their launch pads to the endpoints. We then call Qiskit Metal to draw out these components and feedlines. Since Qiskit Metal currently cannot draw Josephson junctions, we also prepare junction designs with different sizes to be imported by the software. We export our nearly-completed SQP layout and import it into L-Edit [84], an electronic design automation (EDA) software by Mentor Graphics. Finally, we draw flux holes, the bump and UBM pattern, air-tunnels, and possibly TSVs within L-Edit and export the completed SQP layout as a GDSII file.

Figure 4.8 is the workflow diagram of the layout drawing for one of our flip-chip 25-qubit SQPs. As a note, such workflow can be done in principle within only one EDA software. A quantum-specific EDA software is well expected to simplify such a process, preferably with design rule checking (DRC) functionality, for designing large-scale SQPs in the future.

4.5 Design verification

Once there is a completed SQP layout design, it proceeds to the design verification step. During design verification, the first thing is to check that the design satisfies all the constraints imposed during previous design steps. The second is to verify that the layout design is generated correctly according to the designer's expectations. The third is to confirm whether the layout can be identically produced after fabrication. Similar to

traditional semiconductor chip design, this verification step is completed by a group of reviewers formed by design and fabrication experts. After a round of design review session, if no modification is needed for the design and there is no issue with the fabrication feasibility, the GDSII file of the SQP's layout is signed off as final and sent for fabrication.

One of the focus points during the design verification is the correctness of layout generation. There are typical export mistakes such as forgetting to mirror flip the Q-chip design for flip-chip bonding. What's more, when using parameterized design generation tools such as Qiskit Metal, the exported layout may have unexpected outcomes. For example, the rotation of the junction leads for adapting to horizontal and vertical couplers may result in a 1 nm shift of the junction lead width, possibly caused by the handling of the rotation of small objects within the tool.

In our experience, most of the major design revisions are due to issues discovered during the checking of the design's fabrication feasibility. For example, all the structures on the layout should be larger than the minimum geometry that the lithography machine can write on the resist, and be insensitive to the geometric variation due to uncertainty during resist development and metal film etching. What's more, as operators, processes, and tools related to fabrication keep changing and updating, fabrication feasibility criteria will also keep changing and need to be learned. It is possible that an old SQP layout is not feasible under the current fabrication criteria, or it can be done more precisely such that the components' designed parameters can be further optimized. Therefore, a design verification step is recommended whenever a new round of SQP fabrication is planned.

The issues discovered during the verification step can be distilled into rules for future layout design. Some of the rules may be able to be coded into the layout generation program as DRC functionality that can be performed automatically after layout design, similar to the function within traditional EDA software for semiconductor chip design.

4.6 Performance characterization after fabrication

After the fabrication of the SQP, we need to characterize the performance of the obtained processor. To connect the feedlines within the SQP to the outside control and measurement equipments, a PCB is used to fan out the SQP's I/O ports to the cables inside the cryogenic fridge through a package that protects the SQP and the PCB from the outside environment, as shown in Appendix A.

The performance of an SQP can be divided into three levels. *The first level* cares about whether the component parameters of this processor align with the target parameters. This involves qubit frequencies, anharmonicities, qubit-coupler coupling rate, etc, as shown in Table 4.1. *The second level* contains parameters that can only be roughly estimated before fabrication, such as qubit coherence times affected by defects related to fabrication processes and materials. *The third level* is to evaluate the whole SQP by running benchmarking quantum algorithms, such as calculating the cross-entropy fidelity [4].

Figure 4.9 shows one of our fabricated flip-chip integrated SQPs' performance at the first and second levels.

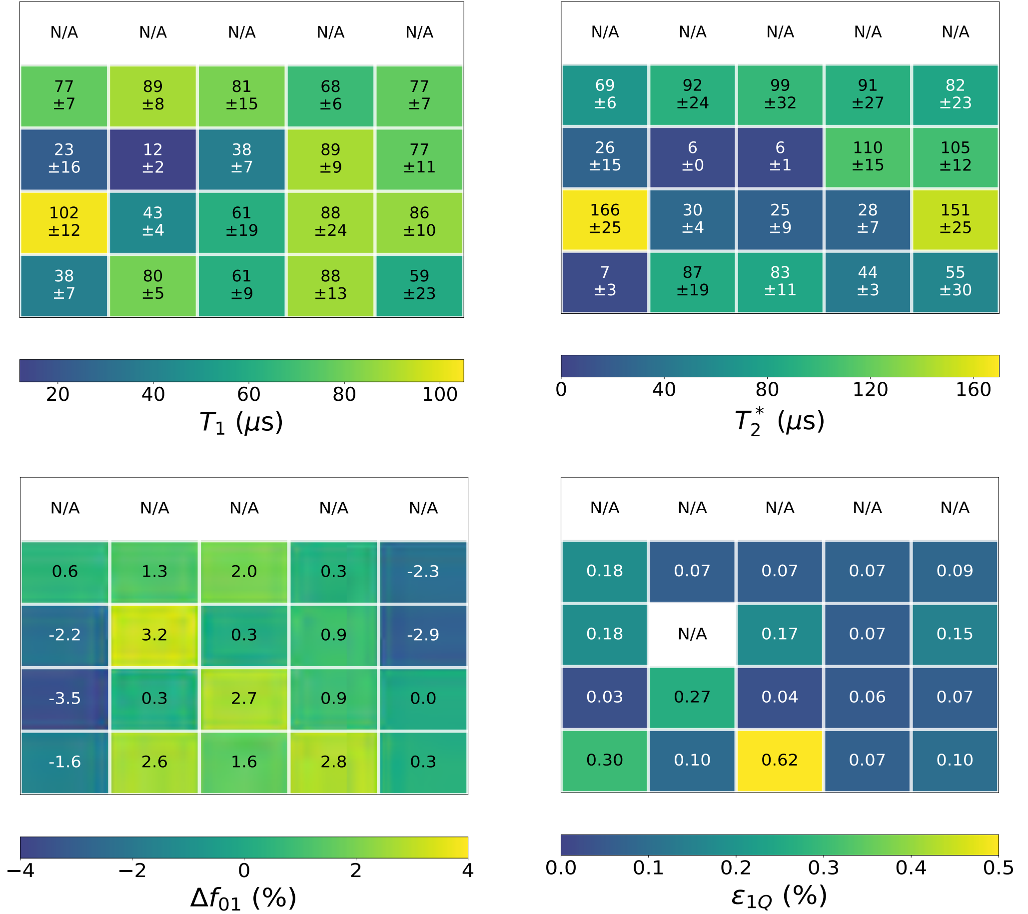


Figure 4.9: Diagrams showing 25-qubit SQP's several performance parameters at first level (qubit frequency deviation from design Δf_{01}) and second levels (qubit coherence T_1, T_2^* [7], single-qubit gate error rate ϵ_{1Q} [63]). The first row of qubits (shown by N/A) were not measured due to limited measurement capacity. Measuring ϵ_{1Q} of one of the qubits was not possible due to readout failure.

4.7 Discussion on scaling up SQPs

The implementation of quantum error-correcting codes is necessary for the realization of fault-tolerant quantum computation. With fixed connectivity between qubits, like in our flip-chip integrated SQPs, quantum error-correcting codes predict that the more qubits we have on the SQP, the better coherence the abstracted logical qubit can obtain, as long as the physical qubit error rates across the SQP are below a certain threshold [71]. In addition, the smaller the physical error, the smaller the number of qubits needed to constitute an error-free logical qubit [85]. For example, using 2D array surface code, an order of 100 qubits is needed if the physical error rate is at 1% of the threshold value, compared to an order of 1000 qubits if the physical error rate is at 10% of the threshold value, for keeping the error rate of a logical qubit below 10^{-10} [70]. Without changing the qubit architecture, our flip-chip integrated SQPs could evolve into a processor of a fault-tolerant quantum computer if we increase the number of qubits on it and at the same time keep the error rate of these qubits smaller than the threshold.

However, scaling up an SQP is not a straightforward task. Apart from structural considerations during the architectural design, all other aspects of the SQP design will encounter obstacles purely due to the large number of qubits.

At the chip specification stage, the number of I/O ports and the connected cables within the cryogenic fridge increase linearly with the qubit number. The first thing we need to do, even before the chip design starts, is to make sure that we have enough equipment to host and characterize such a scaled-up SQP. Simply expanding the number of equipment currently in use will probably be too costly and bulky when the qubit number approaches 1000. Solutions based on cryo-CMOS [86], AQFP logic [54], and SQUID-embedded CPW resonator [55] are receiving more and more attention for mitigating this problem.

On the other hand, if the physical footprint of a unit cell within the SQP doesn't change, the increase in qubit number also means an increase in the size of the chip. Taking our flip-chip integrated SQP's 2 mm qubit-qubit distance as an example, the current largest wafer has a size of 300 mm (12 inches), which means it can host less than $150 \times 150 = 22{,}500$ qubits at maximum. This number is still smaller than the requirement for applying Shor's algorithm practically [24]. However, the larger the chip size, the higher the possibility that defects exist on the chip. The yield of a defect-free SQP will diminish as the size of the chip increases well before we have to use a 12-inch wafer to fabricate. Thus, methods of shrinking the size of the qubits and qubit-qubit distance while not sacrificing the qubit's gate performance are on demand, and some early efforts are reported [52, 58, 87]. In addition, the sizes of the PCB, the package of the chip, and the cryogenic fridge also need to keep up with the size of the SQP, which has their own challenges in maintaining the same performance as before.

At the parameter design stage, if we ignore the interactions beyond a fixed number of qubit neighbors, the number of design parameters needed to be designed will increase linearly with the scale of the SQP. When the qubit number exceeds 100, manually optimizing and checking parameters that satisfy various constraints becomes too cumbersome; thus, an automatic program is needed to assign and verify these target parameters.

Layout design of a large SQP is the most resource-consuming part of the design process. Parameterized design generation tools have largely reduced the manual drawing

of the layout, but still have places that can be improved, such as DRC and auto-routing functionality. The hard part of the layout design is to meet the components' target parameters, which always needs 3D EM simulations as the final step to relate the component's layout to its parameters. If a component's footprint is large, such as a common Purcell filter, the simulation resource it needs will dramatically increase. Therefore, methods to reduce simulation resource for large-footprint components are useful, as the one that will be discussed in Chapter 5. Such large simulations are used not only to determine the component's layout parameters, but also to understand this large component's influence on other components across the SQP. Another kind of simulation concerns the crosstalk between components, which is a daunting task if the number of components included increases. Experiments as in Paper E that measure crosstalk across the SQP are thus necessary to get a sense of how large a crosstalk simulation we should aim for during the layout design.

Similar to the parameter design, the design verification becomes time-consuming when the qubit number is larger than 100. A program to check parameter constraints is expected, and geometry recognition software can play the role of checking the correctness of design layouts. For fabrication feasibility checking, the successful fabrication of a unit cell largely guarantees the feasibility of fabrication of the whole SQP; thus, in most situations, reviewing key areas of the SQP is sufficient.

For the performance characterization of an SQP, even obtaining the basic parameters of qubits becomes cumbersome with a large number of qubits. The usage of the automatic calibration program is required [4] before implementing optimal control techniques for each qubit and running large-scale quantum algorithms.

In summary, scaling up an SQP to become fault-tolerant and be able to solve practical problems is not a straightforward task. We need to keep improving the SQP's overall performance while at the same time decreasing the resource cost per qubit. This task is challenging, but also very exciting and can have substantial rewards.

CHAPTER V

Fast design of coplanar-waveguide (CPW) resonators in flip-chip architecture

To achieve desired qubit control and readout performance, the resonant frequency of the readout resonator and its coupling quality factor to the readout line must meet the target design values. The conventional simulation approach is to vary multiple layout parameters of the resonator's geometry to achieve the desired frequency and coupling quality factor values. However, this approach can be time-consuming for larger devices, if computationally-intensive 3D EM simulation of the entire layout design is used.

This chapter presents a method for predicting the resonant frequency and coupling quality factor of the CPW resonators in the flip-chip architecture. The method makes use of the 2D cross-sections of the resonator, which helps in speeding up the design process and reducing the cost of designing multi-qubit flip-chip-integrated SQPs. With the aid of this fast method, a design strategy for interchip-spacing-insensitive resonators in the flip-chip architecture is proposed, utilizing the opposite responses of the resonant frequency to the interchip spacing when the resonator faces the metal ground plane or the dielectric substrate of the opposing chip. This chapter is based on Paper D with more details.

5.1 Analytical calculation

Typically, a quarter-wave ($\lambda/4$) CPW resonator is coupled capacitively to each qubit in the SQP to enable qubit readout. Figure 5.1(a) shows a conventional design of such a readout resonator. The resonant frequency of this resonator can be calculated by [68]

$$f_r = \frac{1}{4l_{tot}\sqrt{(L_l^g + L_l^k) \cdot C_l}}, \quad (5.1)$$

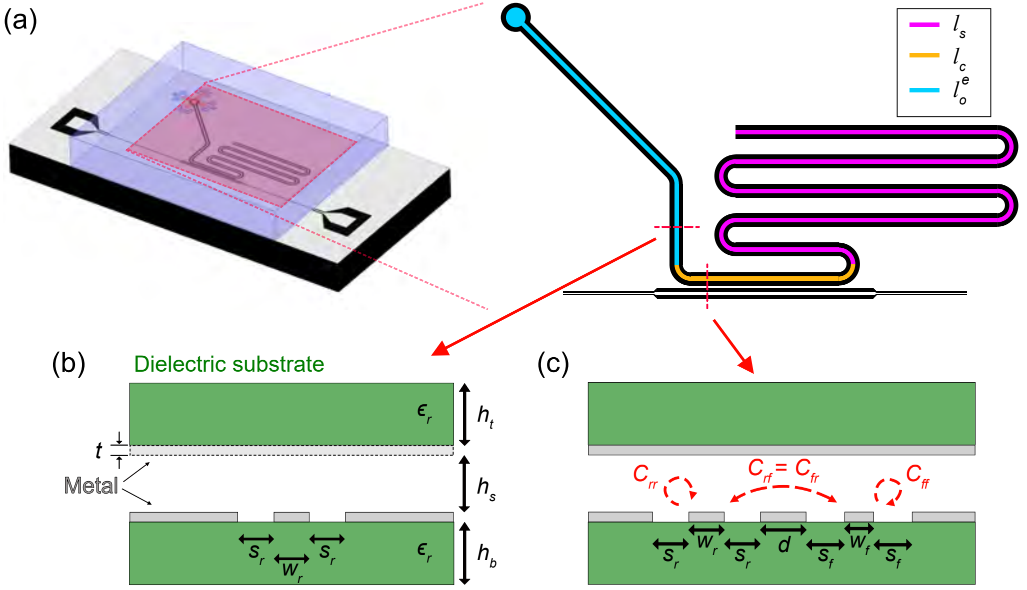


Figure 5.1: (a) A typical $\lambda/4$ CPW readout resonator model in our flip-chip integrated SQP. The zoom-in shows the layout of the resonator, whose body can be divided into three parts (open l_o^e, coupling l_c and short l_s). (b) Cross section of the resonator. Its CPW line can either face the metal ground plane of the top chip, or the dielectric substrate directly. (c) Cross section of the resonator's coupling part with the readout line. The capacitance between the resonator's and the readout line's center conductors (C_{rf}=C_{fr}) and their self-capacitances (C_{rr}, C_{ff}) are simulated to obtain the coupling quality factor Q_c to the readout line.

where l_{tot} is the resonator's total length, including the effective lengths resulting from discontinuities at both ends of the CPW line. L_l^g and L_l^k are the geometric and kinetic inductance per unit length, respectively, and C_l is the capacitance per unit length of the CPW line.

We can determine the L_l^g and C_l of a resonator by analyzing its 2D cross-section. This can be done either through analytical calculation or EM simulation. We assume that the thin metal films shown in Figure 5.1(b)-(c) are perfect conductors with no EM field inside them. Combined with the assumption of zero thickness ($t = 0$), we can simplify the analytical calculation. Simple conformal mapping techniques [88] can then be used to obtain the values of L_l^g and C_l.

For planar CPW lines in the absence of the top chip, i.e. in the single-chip architecture, L_l^g and C_l can be obtained by [89]

$$L_l^g = \frac{\mu_0}{4} \cdot \frac{K(k_1')}{K(k_1)}, \tag{5.2}$$

$$C_l = 2\varepsilon_0(\varepsilon_r - 1) \cdot \frac{K(k_2)}{K(k_2')}, \tag{5.3}$$

where μ_0 is the vacuum permeability, ε_0 is the vacuum permittivity, ε_r is the relative permittivity of the dielectric substrate, and $K(k)$ is the complete elliptic integral of the first kind with modules given as

$$k_1 = \frac{w_r}{w_r + 2s_r}, \tag{5.4}$$

$$k_2 = \sinh\left[\frac{\pi w_r}{4h_b}\right] / \sinh\left[\frac{\pi(w_r + 2s_r)}{4h_b}\right], \tag{5.5}$$

$$k_i' = \sqrt{1 - k_i^2}\ (i = 1, 2), \tag{5.6}$$

where w_r and s_r are the center conductor and gap width of the CPW line, and h_b is the substrate thickness of the bottom chip.

As shown in Figure 5.1(b), for CPW resonators facing the metal ground plane of the top chip, its $L_l^{g,\mathrm{m}}$ and C_l^{m} values are calculated as [90, 91]

$$L_l^{g,\mathrm{m}} = \frac{\mu_0}{2}\left[\frac{K(k_s)}{K(k_s')} + \frac{K(k_1)}{K(k_1')}\right]^{-1}, \tag{5.7}$$

$$C_l^{\mathrm{m}} = 2\varepsilon_0 \frac{K(k_s)}{K(k_s')} + 2\varepsilon_0\left[\frac{K(k_1)}{K(k_1')} + (\varepsilon_r - 1)\frac{K(k_2)}{K(k_2')}\right], \tag{5.8}$$

with

$$k_s = \tanh\left[\frac{\pi w_r}{4h_s}\right] / \tanh\left[\frac{\pi(w_r + 2s_r)}{4h_s}\right]. \tag{5.9}$$

$$k_s' = \sqrt{1 - k_s^2}, \tag{5.10}$$

where h_s is the interchip spacing of the two chips.

For CPW resonators facing the dielectric substrate of the top chip, their $L_l^{g,\mathrm{d}}$ and C_l^{d} are obtained by [92]

$$L_l^{g,\mathrm{d}} = \frac{\mu_0}{4}\frac{K(k_1')}{K(k_1)}, \tag{5.11}$$

$$C_l^{\mathrm{d}} = \frac{2\varepsilon_0}{\left(\varepsilon_r \frac{K(k_1)}{K(k_1')}\right)^{-1} + \left(\frac{\varepsilon_r}{\varepsilon_r - 1}\frac{K(k_s)}{K(k_s')}\right)^{-1}} + 2\varepsilon_0\left[\frac{K(k_1)}{K(k_1')} + (\varepsilon_r - 1)\frac{K(k_2)}{K(k_2')}\right]. \tag{5.12}$$

A detailed derivation of L_l^g and C_l for all three cases above is illustrated in Appendix B.

The kinetic inductance L_l^k arises from the inertial of charge carriers. For superconductors, L_l^k can be thought of as coming from Cooper pairs in the two-fluid model [93, 94]. To calculate L_l^k in Equation 5.1, we only need to look at the supercurrent distribution inside the metal thin films. The calculation is shown by [95]

$$L_l^k = \frac{\mu_0 \lambda_m^2}{|I|^2} \cdot \int J_z^2\, dS, \tag{5.13}$$

where J_z is the supercurrent density in the direction of the current flow (normal to the plane), and the surface integral is over the cross-section of the thin film only. For a flip-chip geometry, the integral also includes the top chip's metal ground plane. Furthermore, λ_m is the magnetic penetration depth of the superconductor, and I is the total current injected into the CPW's center conductor.

There are analytical equations of L_l^k for planar CPW lines [95, 96] based on the assumption of homogeneous supercurrent density within the penetration depth. However, within the flip-chip architecture, the supercurrent distribution may be significantly affected by the metal film on the top chip. Since there is no available analytical equation for CPW in such a situation, we choose to simulate J_z to obtain L_l^k. The detailed simulation method is explained in Paper D and Reference [97].

After obtaining J_z, the magnetic penetration depth λ_m must be known to calculate L_l^k. However, for very thin metal film ($t < 10\lambda_m$), λ_m is a sensitive value that depends on the actual metal film quality and thickness [98]. It is then necessary to determine the λ_m uniquely for our lab's particular fabrication process. By comparing our fabricated planar resonators' resonant frequencies with the simulated values that omit the effect of L_l^k, we can extract that our aluminum metal thin film has the magnetic penetration depth of $\lambda_m = 83$ nm, which is larger than that of bulk value [99, 100] and our film is in dirty limit.

For a resonator that is used for qubit readout, its coupling quality factor Q_c to the readout line is the second parameter that needs to be obtained. Here, we focus on the Q_c when the resonator, made of CPW lines, is coupling to another CPW line (readout line) by letting a section of the resonator be parallel to the readout line (Figure 5.1(a) and (c)). We deal with the case when the characteristic impedance of the readout line matches the I/O port (50 Ω). The resulting equation is [101]

$$\frac{1}{Q_c} = \frac{2\kappa^2 \sin^2\theta}{\pi(2p-1)}. \tag{5.14}$$

The resonator's frequency shift due to the coupling, δf_r^c, is

$$\delta f_r^c = -\frac{c_l \sin\theta}{2\pi l_{tot}} \cdot \left[\frac{\kappa^2(2\cos\psi + \cos\theta)}{2} + \frac{(Z_2 - Z_r)\cos\psi}{Z_r}\right], \tag{5.15}$$

with

$$\kappa = -C_{rf}/\sqrt{C_{rr}C_{ff}}, \tag{5.16}$$

$$c_l = f_r \cdot 4l_{tot}, \tag{5.17}$$

$$Z_2 = 1/\left(c_l C_{ff}\sqrt{1-\kappa^2}\right), \tag{5.18}$$

$$\theta = 2\pi l_c/(4l_{tot}), \tag{5.19}$$

$$\psi = 2\pi(l_c + 2l_o^e)/(4l_{tot}), \tag{5.20}$$

where f_r is the resonator's bare frequency in the absence of the readout line, and we take the integer number $p = 1$ for its fundamental resonance. As shown in Figure 5.1(a), l_c

is the length of the coupling part between the resonator and the readout line, including two additional 90-degree arcs at both ends of the coupling part to take into account the spurious coupling [101]; l_o^e is the effective length of the open part of the resonator, including the effective length of the coupling structure to the qubit; Z_r is the characteristic impedance of the resonator's CPW line, and c_l is the speed of light within the resonator's CPW cross-section. The coupling capacitance ratio κ and the impedance Z_2 of the resonator's coupling part are calculated from the capacitance between the resonator's and the readout line's center conductors C_{rf} and their self capacitances (C_{rr}, C_{ff}) as illustrated in Figure 5.1(c).

The fact that the coupling capacitance ratio κ and the impedance Z_2 can be represented by capacitance alone is due to the assumption that the CPW line works in the quasi-TEM mode [68].

5.2 Comparison to 3D simulations

5.2.1 Resonant frequency f_r

To show the analytical calculations by using conformal mapping techniques and 2D EM simulations are in good agreement with 3D EM simulations, Figure 5.2(a)-(b) compares the resonator's resonant frequencies obtained by these methods at different interchip spacings h_s when it is facing the metal ground plane of the top chip. f_r^{conf} is obtained by using $L_l^{g,\mathrm{m}}$ and C_l^{m} derived from Equation 5.7 and 5.8. f_r^{sim2D} is obtained by directly simulating $L_l^{g,\mathrm{m}}$ and C_l^{m} in a 2D FEM simulation. f_r^{sim3D} is found by directly modeling the resonator in a 3D FEM simulation. We can see that the deviation between various methods is less than 2% when h_s is larger than 3 μm. As a note, we didn't take L_l^k into account in this case as the modeled metal thin films within the 3D simulations are set to be perfect conductors. In Paper D we show the comparison of the resonator's resonant frequencies obtained by various methods including measured values within a multi-qubit superconducting quantum processor, in which L_l^k is considered.

Figure 5.2(c)-(d) shows the comparison between 3D and 2D methods to obtain the resonator's resonant frequency when the resonator faces the dielectric substrate of the top chip. $f_r^{\mathrm{conf}'}$ is obtained by using $L_l^{g,\mathrm{d}}$ and C_l^{d} from Equation 5.11 and 5.12, while $f_r^{\mathrm{sim2D}'}$ is obtained by directly simulating $L_l^{g,\mathrm{d}}$ and C_l^{d} in a 2D FEM simulation and $f_r^{\mathrm{sim3D}'}$ is obtained by a 3D FEM simulation. The deviation between the 2D and 3D methods is below 5% when h_s is larger than 2 μm.

5.2.2 Coupling quality factor Q_c

Figure 5.3 compares 2D and 3D methods for obtaining the resonator's Q_c and calculated resonator's frequency shift δf_r^c at different h_s. Q_c^{cal2D} and $Q_c^{\mathrm{cal2D}'}$ are obtained by Equation 5.14 plus the 2D EM simulated capacitances (C_{rf}, C_{rr}, C_{ff}) when the resonator faces either the metal ground plane or the dielectric substrate of the top chip, respectively. Q_c^{sim3D} and $Q_c^{\mathrm{sim3D}'}$ are obtained by direct 3D FEM simulations also in two cases.

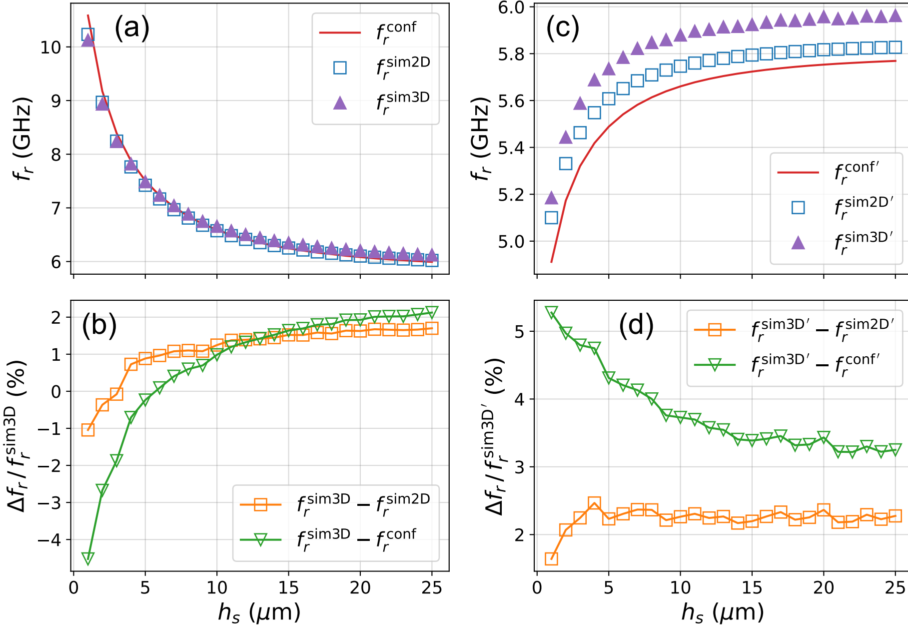


Figure 5.2: Resonant frequency comparison when the CPW resonators face either the metal ground plane or the dielectric substrate of the top chip. (a)&(c) Resonator frequencies obtained from conformal mapping calculation (metal on top: f_r^{conf}, dielectric on top: $f_r^{\mathrm{conf}'}$), 2D FEM simulation (metal on top: f_r^{sim2D}, dielectric on top: $f_r^{\mathrm{sim2D}'}$), and 3D FEM simulation (metal on top: f_r^{sim3D}, dielectric on top: $f_r^{\mathrm{sim3D}'}$) under different interchip spacings h_s. It is noticeable that under these two scenarios the f_r exhibits opposite trends with the change of h_s. (b)&(d) Frequency difference Δf_r between 3D and 2D FEM simulation, and between 3D FEM simulation and conformal mapping calculation.

5.3 Discussion on computational resources consumption

Most of the computational resources are consumed during the layout design stage of the SQP. After determining the target parameters of the components within a multi-qubit SQP, EM simulations are necessary to figure out the corresponding layout design of each component on the processor.

To determine the layout design of all components on the processor, two types of EM simulations are conducted. *The first type* involves a rough sweep of the layout parameters of a specific component design. For instance, by sweeping the meandering length of the resonator with large steps, we can establish the relationship between the resonator's resonant frequency and its meandering length while other parameters are fixed. This helps us quickly estimate the resonators' meandering length with different target resonant frequencies. The reason for this coarse simulation is that a target parameter of the component is generally influenced by several layout parameters. If we were to

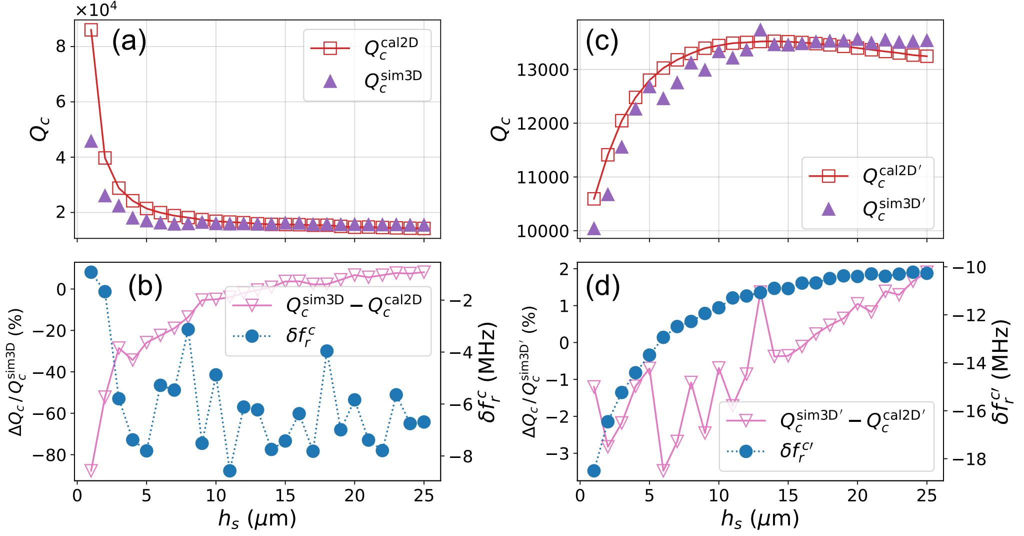


Figure 5.3: Coupling quality factor comparison when the CPW resonator faces either the metal ground plane or the dielectric substrate of the top chip. (a)&(c) Resonator coupling quality factor obtained from 2D cross-sections (metal on top: Q_c^{cal2D}, dielectric on top: $Q_c^{\text{cal2D}'}$) and 3D FEM simulation (metal on top: Q_c^{sim3D}, dielectric on top: $Q_c^{\text{sim3D}'}$). (b)&(d) their difference, together with the coupling-induced frequency shift (metal on top: δf_r^c, dielectric on top: $\delta f_r^{c'}$) under different h_s. δf_r^c ($\delta f_r^{c'}$) is calculated using f_r^{sim2D} ($f_r^{\text{sim2D}'}$) as the bare resonator frequency.

conduct detailed sweeping simulations involving all the parameters in a wide range, the total number of simulations would become large, while most of the runs are actually unnecessary.

The second type of EM simulation involves determining the layout design of each component separately. This is done by performing a fine sweep of layout parameters around the component's target parameter. At this time, additional structures around the component can also be included in the simulations, which helps to capture the side influence of these structures, for example, the resonator's resonant frequency shift due to coupling to the readout line.

We can then estimate the computational resources necessary to translate a processor's parameter design into its layout design. We can calculate the required computational resources (central processing unit (CPU) time × memory) as follows:

$$\text{Total resource needed} = \text{Number of simulations} \times \text{Resources per simulation run},$$

where

$$\begin{aligned}\text{Number of simulations} = {} & \text{Number of parameters} \times \\ & (\text{Coarse sweep steps} + \text{Number of components} \times \text{Fine sweep steps}).\end{aligned}$$

The number of simulations required for a specific component design within a processor is usually fixed. Therefore, the total resources needed for the layout design stage can be

reduced by lowering the necessary resources per simulation run. For example, as shown in Paper D, a single 3D EM simulation for a readout resonator in the flip-chip architecture requires an average of 32 hours of CPU time and 58 GB of memory allocation. On the other hand, using 2D EM simulations, the average CPU time spent is only 1.7 minutes, and the maximum memory allocation is 64 MB per simulation run. Compared to 3D EM simulations, the computational resource (CPU time × memory) is reduced by a factor of 1000 at each simulation. This means that if most of our simulations are in 2D rather than 3D, the SQP's layout design stage can be carried out on a computer with smaller memory, using fewer CPUs, or requiring shorter simulation time.

5.4 Making f_r insensitive to interchip spacing

Figure 5.2 clearly shows that the change in the resonator's resonant frequency f_r with the change of interchip spacing h_s behaves differently depending on whether it faces the metal ground plane or the dielectric substrate of the top chip. Therefore, if we can combine these opposite trends within one resonator by an optimized cutout area on the top chip's metal ground plane which covers the resonator meandering part (see Figure 5.4(a)), we can in principle suppress its sensitivity to the variation of h_s around the target value, e.g., $h_s = 8$ µm.

Using Equations 5.7-5.12, we can derive a ratio γ between the area of the metal part and the dielectric part that the resonator is facing on the top chip such that the variation of the resonator's f_r is minimized. Alternatively, we can also use direct 3D simulations to obtain this optimum ratio. It can be noticed that in Figure 5.4(b) there is a discrepancy between the calculated and simulated ratio due to the meandering structure of the resonator. We recommend using the calculated ratio as the starting point and then sweeping in 3D simulation to determine the final interchip-spacing-insensitive resonator design.

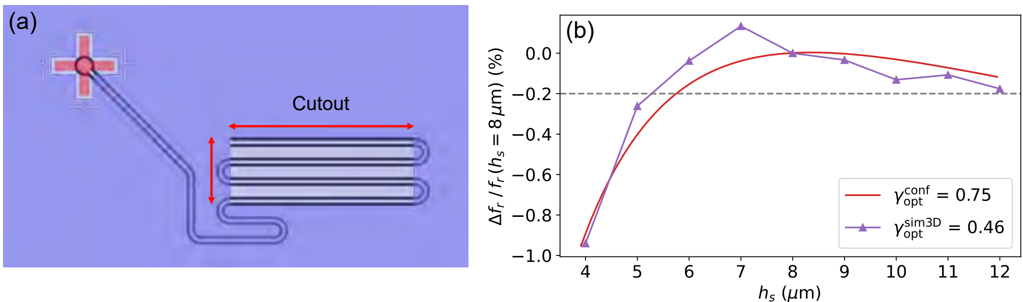


Figure 5.4: (a) Top view of a $\lambda/4$ resonator facing a partial cut-out of the metal ground plane on the opposing chip. (b) f_r's variation around $h_s = 8$ µm for $\gamma = \gamma_{\text{opt}}$, obtained from conformal mapping calculation and 3D EM simulation, respectively.

5.5 Q-factor of the resonator's cross-section in flip-chip architecture

In Chapter 3 we have used Equation 3.1 and 2D EM simulations to compare the Q-factors of a CPW line's cross-section when the top chip is present or not. Here we can do such a comparison again but in this case, the CPW line either faces a metal ground plane or a bare dielectric substrate on top.

Figure 5.5 shows the Q-factors of the resonator cross-section when its CPW line is facing either the metal ground plane or the dielectric substrate of the top chip at different interchip spacings h_s. As h_s decreases, there is a small increase of Q-factor until h_s is below a certain threshold. We also notice that the Q-factor is slightly lower when the CPW line is facing the dielectric substrate. The Q-factor drops significantly at small h_s when the CPW line faces the metal ground plane, which is caused by the increased E-field strength inside the metal–air dielectric interfaces when the two chips are very close to each other.

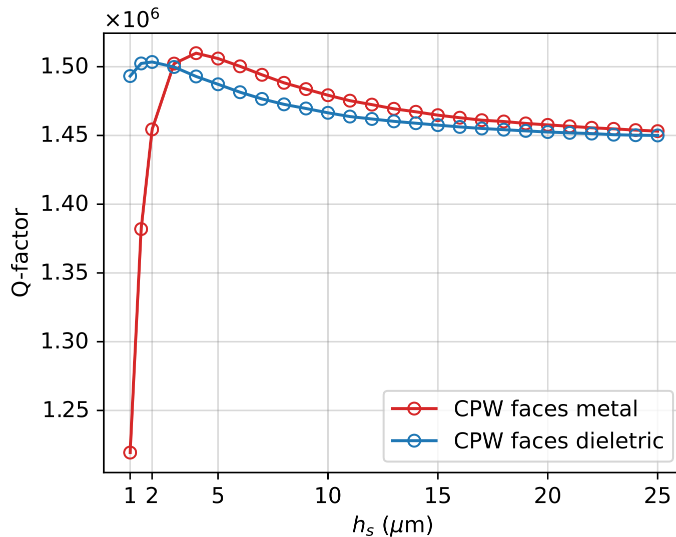


Figure 5.5: Cross-sectional Q-factor when the resonator's CPW line faces either the metal ground plane or the dielectric substrate of the top chip at different interchip spacings h_s.

CHAPTER VI

Summary and contributions

6.1 Summary

To realize fault-tolerant quantum computers, and be able to run practical quantum algorithms, current quantum computers need to be scaled up. However, scaling up quantum computers is not an easy task. This thesis summarizes our work on improving the scalability of superconducting quantum processors by adopting 3D integration technologies used in the semiconductor industry, resulting in flip-chip integrated superconducting quantum processors. We demonstrated a processor whose constituents (qubits, readout resonators, couplers, Purcell filters, etc.) maintain good performance when integrated in a flip-chip architecture, compared to our previous single-chip, planar architecture. Furthermore, to demonstrate the improved scalability of this new architecture, we developed a design workflow for multi-qubit superconducting quantum processors and designed, fabricated, and characterized a 25-qubit flip-chip integrated processor. In addition, to speed up our design workflow, we showed that we can greatly reduce the simulation time and resources of designing the readout resonators within the processors by switching most of the simulations from 3D to 2D.

Efforts are continually being put into this scaling-up task, in our group and around the world. We can envision that additional 3D integration technologies will be introduced for superconducting quantum processors, accompanied by new designs of components, with new materials and fabrication techniques. The design workflow will be further optimized with new tools and processes. And control and supporting equipment will also evolve with the scaled-up processors.

To conclude, superconducting quantum computing has the potential to proceed and enter the realm of applications, and I am proud of having contributed to this goal.

6.2 Contributions

In this section, I list my contributions to the publications appended at the end of the thesis, and I also list my additional works presented in the thesis:

- **Paper A**: I am the co-1st author of the paper. Together with Sandoko Kosen, I designed, simulated, and characterized basic components in superconducting quantum processors (resonators, qubits, and couplers) in flip-chip devices, and also printed circuit boards and sample holders. I also participated in the device fabrication and the writing of the paper.
- **Paper B**: I contributed to the characterization of the internal quality factor of the resonators on chips in the presence of through-silicon vias.
- **Paper C**: I contributed to the design, simulation, and characterization of the flip-chip superconducting quantum processor used in the paper and wrote the appendix of the paper.
- **Paper D**: I conducted the analytical calculations, the numerical simulations, and wrote the paper. I developed and verified the idea of interchip-spacing insensitive flip-chip resonator design. I contributed to the design, fabrication, and characterization of the multi-qubit flip-chip-integrated superconducting quantum processor used in the paper.
- **Paper E**: I contributed to the design, fabrication, and characterization of the multi-qubit flip-chip-integrated superconducting quantum processor used in the paper.
- **In thesis**: I designed, simulated, and characterized the 25-qubit flip-chip-integrated superconducting quantum processor that hosts common Purcell filters, and analyzed the measured data presented in the thesis.

APPENDIX I

Packaging of an SQP

After the fabrication of an SQP, another step must be taken from having an SQP at hand to characterizing its performance. This step is called chip packaging. This appendix illustrates a typical packaging procedure at our lab.

After the flip-chip bonding of Q-chip and C-chip, a flip-chip integrated SQP is ready for characterization. We put the processor at the center of a corresponding PCB, and both the SQP and PCB are sitting on top of a sample holder. The PCB is fixed to the sample holder by screws, while the SQP is glued on top of it.

After several hours of waiting for the glue to dry, we then start connecting the SQP with the PCB by wire bonding at four edges of the SQP. We use an automatic wire-bonding machine to complete this task. Figure A.1(a) shows the machine while it is conducting wire bonding according to pre-set wire locations, and Figure A.1(b) shows the sample holder with the PCB and SQP after wire bonding.

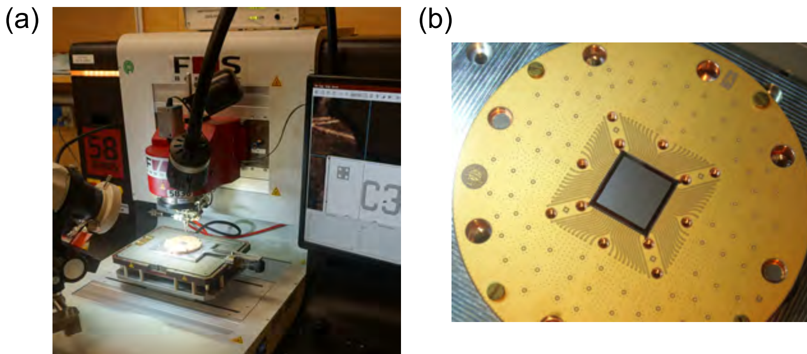


Figure A.1: (a) The automatic wire-bonding machine in our lab. (b) The sample holder (copper) that hosts an 80-port PCB (gold-plated) and a 25-qubit flip-chip integrated SQP at its center.

After the SQP is wire-bonded to the PCB, and a copper cap is placed on top of the SQP to provide mechanical protection and E-field shielding (Figure A.2(a)), we can then put the whole sample holder into the mixing chamber of a cryogenic fridge, as shown in Figure A.2(b). Within the fridge, the cables are wired through the mixing chamber to the outside of the fridge through several temperature stages, accompanying signal filtering and amplification, as shown in Figure A.2(c) and Figure A.3.

Then we close the fridge and cool it down until the mixing chamber temperature is around 10 mK. The SQP is now ready for characterization by using electronic control equipment beside the fridge at room temperature to send signals and receive responses from the SQP.

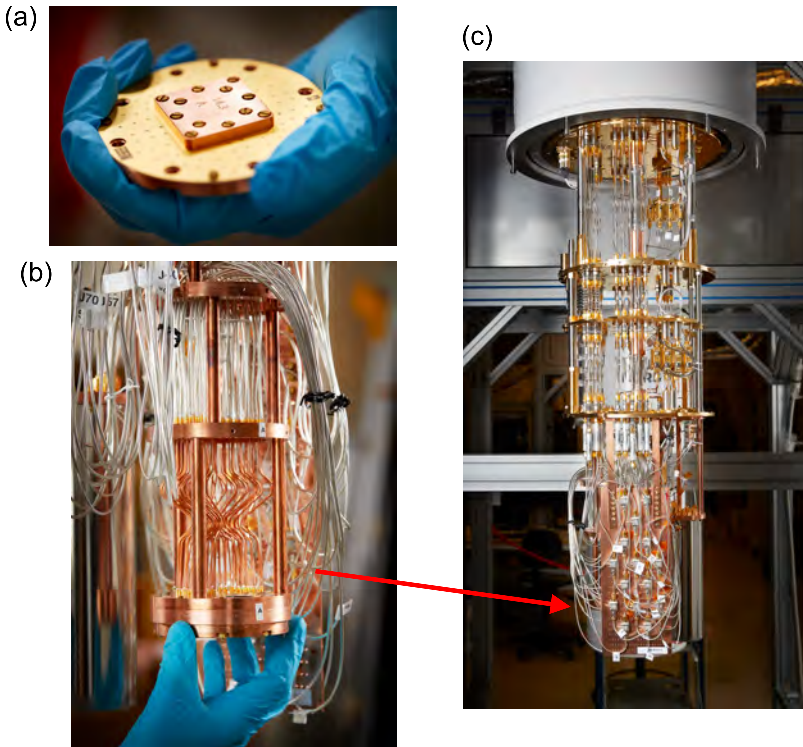


Figure A.2: (a) The SQP is fully packaged after a copper cap is put on top of the SQP. (b) The sample holder is then attached to the I/O cable tower from the bottom. (c) The packaged SQP will be sitting in the mixing chamber stage during the characterization, which is at the bottom of the cryogenic fridge.

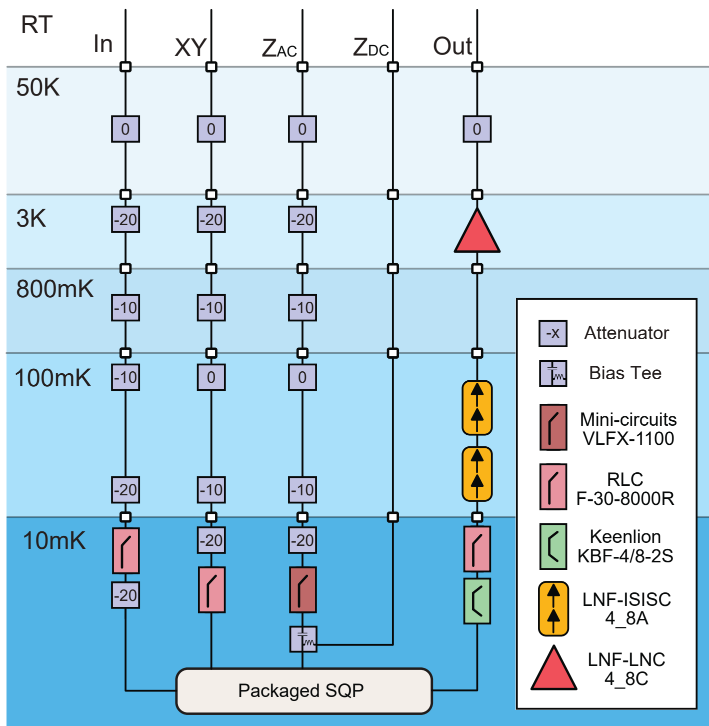


Figure A.3: The wiring diagram within the cryogenic fridge for qubit's control and readout.

APPENDIX II

Derivation of L_l^g, C_l of a CPW line

In this appendix, we will use conformal mapping techniques to derive the geometric inductance L_l^g and capacitance per unit length C_l of the coplanar waveguide's cross-section. We will present the derivations when the CPW line is in planar geometry and flip-chip geometry.

We assume a zero thickness of the metal thin films to simplify our conformal transformation functions so that they can be treated analytically. The basic principle behind the conformal mapping technique is to transform a given transmission line cross-section to simple parallel-plate type like the one shown in Figure B.1, such that L_l^g, C_l are calculated using

$$L_l^g = \mu_0 \frac{H}{W}, \tag{B.1}$$

$$C_l = \varepsilon_r^p \varepsilon_0 \frac{W}{H}, \tag{B.2}$$

where H is the separation between the two plates, W is the width of the plates, and ε_r^p is the relative permittivity of the dielectric medium between two plates after the conformal transformation.

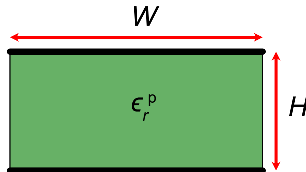


Figure B.1: Cross-section of a simple parallel-plate waveguide. ε_r^p is the relative permittivity of the dielectric medium between two plates.

B.1 Planar CPW line

We first show how to apply the conformal transformation to a planar CPW cross-section having a dielectric substrate with relative permittivity ε_r, and vacuum above the CPW line. Figure B.2(a) shows the cross-section of such a CPW line.

Geometric inductance L_l^g:

We first calculate the geometric inductance per unit length of the cross-section. We replace the substrate with the vacuum since their permeability is the same. Then we put the cross-section into a complex z-plane. Because all the metal layers are in the horizontal direction, they can be put along the Re[z]-axis and the Christoffel-Schwartz transformation [88] is applied to conformally map these metals into a parallel-plate waveguide in the complex w-plane [89]. The transformation function which we use is

$$w(t) = A_1 \int_0^z \frac{dz}{\sqrt{(z - z_B)(z - z_C)(z - z_D)(z - z_E)}} + A_2. \tag{B.3}$$

Here A_1 and A_2 are constants that determine the scaling and translation of the transformed geometry, and z_i $(i = B, C, D, E)$ are the positions of the endpoints of the metals on the Re[z]-axis, as illustrated in Figure B.2(a). Choosing the center of the CPW's center conductor as the zero position of the Re[z]-axis, we have $-z_B = z_E = (w_r + 2s_r)/2$ and $-z_C = z_D = w_r/2$.

The geometry in the z-plane is now conformally mapped to the w-plane. As a result, the CPW center conductor and the ground plane (two infinite points viewed as connected) in the z-plane are transformed into the two plates with equal width and separated in parallel, forming a parallel-plate waveguide with a vacuum between the plates.

The width and the height of this parallel-plate waveguide are calculated as

$$W_b^{\text{vac}} = |w_D - w_C| = |A_1| 2K(k_1), \tag{B.4}$$

$$H_b^{\text{vac}} = |w_E - w_D| = |A_1| K(k_1'), \tag{B.5}$$

where $K(k)$ is the complete elliptic integral of the first kind with modules $k_1 = z_D/z_E = w_r/(w_r + 2s_r)$ and $k_1' = \sqrt{(1 - k_1^2)}$.

The geometric inductance per unit length of this parallel-plate waveguide is then obtained from

$$L_l^g = \mu_0 \frac{H_b^{\text{vac}}}{W_b^{\text{vac}}} = \frac{\mu_0}{2} \frac{K(k_1')}{K(k_1)}. \tag{B.6}$$

Capacitance C_l:

To calculate the capacitance per unit length, we need to consider the substrate of the CPW line. We can view the cross-section (Figure B.2(a)) as two cross-sections in a parallel combination, in which the first has the vacuum below the metal layer, whereas the second has a finite-thickness dielectric substrate with relative permittivity $\varepsilon_r - 1$. We

can then calculate separately the capacitance contributions from the vacuum C_l^{vac} and the substrate C_l^{sub}, and sum over the two to get C_l.

C_l^{vac} can be calculated using the same conformal transformation function (Equation B.3) as in calculating L_l^g. Therefore we have

$$C_l^{\text{vac}} = \varepsilon_0 \frac{W_b^{\text{vac}}}{H_b^{\text{vac}}} = 2\varepsilon_0 \frac{K(k_1)}{K(k_1')}. \tag{B.7}$$

For the cross-section having a finite-thickness substrate with relative permittivity $\varepsilon_r - 1$, we can first do an intermediate transformation such that the substrate becomes infinitely thick to resemble the vacuum case above. We map the cross-section from the z-plane to the t-plane with the function

$$t(z) = \sinh\left[\frac{\pi z}{2h_b}\right]. \tag{B.8}$$

From the t-plane, we repeat the same Christoffel-Schwartz transformation but replace the variable notations in Equation B.3 from z to t. Figure B.2 shows the two consecutive conformal transformations.

Thus we obtain

$$C_l^{\text{sub}} = (\varepsilon_r - 1)\varepsilon_0 \frac{W_b^{\text{sub}}}{H_b^{\text{sub}}} = 2(\varepsilon_r - 1)\varepsilon_0 \frac{K(k_2)}{K(k_2')}, \tag{B.9}$$

with modules

$$\begin{aligned} k_2 = t_D/t_E &= \sinh\left[\frac{\pi z_D}{2h_b}\right] / \sinh\left[\frac{\pi z_E}{2h_b}\right] \\ &= \sinh\left[\frac{\pi w_r}{4h_b}\right] / \sinh\left[\frac{\pi(w_r + 2s_r)}{4h_b}\right], \end{aligned} \tag{B.10}$$

$$k_2' = \sqrt{1 - k_2^2}. \tag{B.11}$$

Therefore, for the planar CPW line's cross-section, we have:

$$L_l^g = \frac{\mu_0}{2} \frac{K(k_1')}{K(k_1)}, \tag{B.12}$$

$$\begin{aligned} C_l &= C_l^{\text{vac}} + C_l^{\text{sub}} \\ &= 2\varepsilon_0 \left[\frac{K(k_1)}{K(k_1')} + (\varepsilon_r - 1)\frac{K(k_2)}{K(k_2')}\right]. \end{aligned} \tag{B.13}$$

B.2 Flip-chip CPW line

For the CPW line in the flip-chip geometry, we assume that the interchip spacing h_s between two chips is large enough, such that the magnetic field around the CPW center conductor is perpendicular to the vacuum-dielectric interfaces at gap areas of the CPW

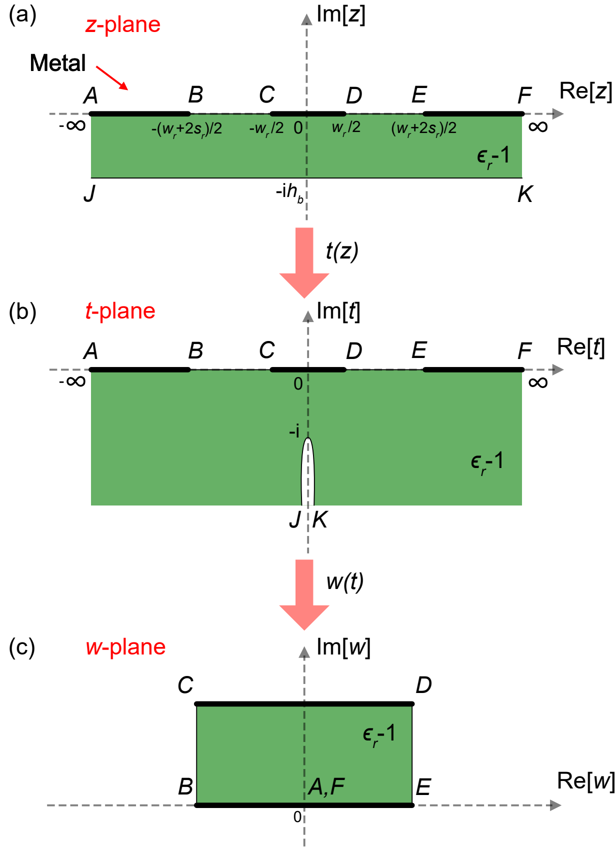


Figure B.2: Conformal transformations for the planar CPW cross-section. (a) Original geometry with substrate relative permittivity of $\varepsilon_r - 1$ after removing the contribution from the vacuum. (b) Intermediate geometry in the t-plane. (c) Final transformation to a parallel-plate waveguide.

line. We can then cover the vacuum-dielectric interfaces with magnetic walls and separate the CPW cross-section into two halves. As shown in Figure B.3, the bottom half is just the same as the planar CPW line, and also its contributions $L_l^{g,b}$ and C_l^b to the total L_l^g, C_l. However, due to the presence of the top chip, the contributions from the top half of the cross-section are different from those of a pure vacuum.

We independently apply conformal transformations to the top half, resulting in a different parallel-plate waveguide with a different width W_t and separation H_t. The contributions of two halves to the total L_l^g, C_l of the CPW cross-section are added as a parallel combination:

$$L_l^g = \left(1/L_l^{g,t} + 1/L_l^{g,b}\right)^{-1}, \tag{B.14}$$

$$C_l = C_l^t + C_l^b, \tag{B.15}$$

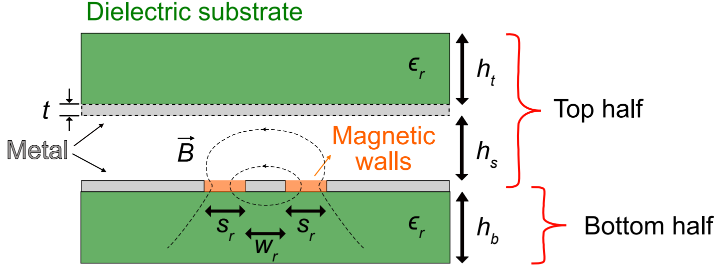


Figure B.3: Cross-section of the CPW line in the flip-chip geometry. The top and bottom halves of the cross-section, separated by the metal surface of the bottom chip, can be independently transformed into two parallel plates using conformal mapping techniques, when the magnetic walls (orange) are placed on the surfaces of the CPW's gap area. We will consider the two scenarios in which the metal ground plane of the top chip is present or not.

where $L_l^{g,b}$ and C_l^b are the same as the planar CPW line case.

B.2.1 Top chip with the ground plane

When the top-half cross-section has the metal ground plane on top of the CPW line, this superconducting metal thin film shields all the electromagnetic fields generated from the CPW line. Therefore, we can view the top chip's substrate as absent and replace it with a vacuum.

To simplify the calculation, we exploit the symmetry of the top-half cross-section and only calculate $L_l^{g,t}$ and C_l^t at the z-plane's real positive part (zero position is at the center of CPW's center conductor). See Figure B.4(a). The resultant two identical parallel-plate waveguides mapped from the real positive and negative parts are also treated as a parallel combination.

We do two consecutive conformal transformations to map the real positive part of the top-half cross-section into a parallel-plate waveguide. We first map the top chip's half of the metal ground plane in the z-plane to the Re[t]-axis and then use the Christoffel-Schwartz transformation to map the geometry to a parallel-plate waveguide in the w-plane [91]. Figure B.4 shows the two conformal transformations.

The first transformation function we use is

$$t(z) = \cosh^2\left[\frac{\pi z}{2h_b}\right]. \tag{B.16}$$

In the z-plane, we define $z_P = 0$, $z_I = ih_s$, $z_D = w_r/2$ and $z_E = (w_r + 2s_r)/2$. In the t-plane these points are mapped into $t_P = 1$, $t_I = 0$, $t_D = \cosh^2[\pi w_r/4h_b]$ and $t_E = \cosh^2[\pi(w_r + 2s_r)/4h_b]$.

The second transformation function we use is

$$w(t) = A_1 F(\varphi, k_s) + A_2, \tag{B.17}$$

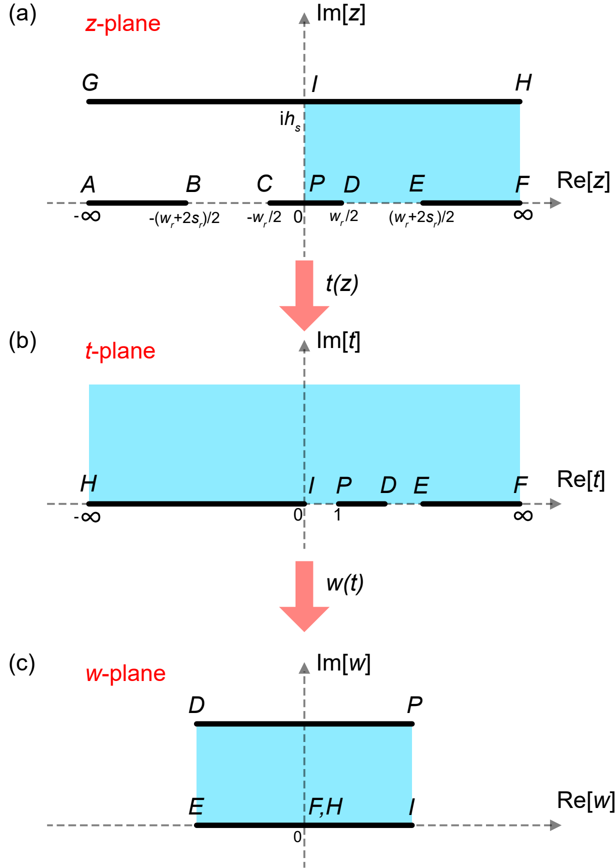


Figure B.4: Conformal transformations for the top half of the CPW cross-section in the flip-chip geometry, where the CPW line is covered by the metal ground plane of the top chip. The transformed region is painted in cyan. (a) Original geometry. (b) Intermediate geometry in the t-plane. (c) Final transformation to a parallel-plate waveguide.

where $F(\varphi, k_s)$ is the elliptic integral of the first kind with

$$F(\varphi, k_s) = \int_0^{\sin\varphi} \frac{d\tau}{\sqrt{(1-k_s^2\tau^2)(1-\tau^2)}}, \tag{B.18}$$

$$\sin\varphi = \sqrt{\frac{(t-t_E)t_D}{(t-t_D)t_E}}, \tag{B.19}$$

$$k_s = \sqrt{\frac{t_E(t_D-t_P)}{t_D(t_E-t_P)}} = \tanh\left[\frac{\pi w_r}{4h_s}\right] / \tanh\left[\frac{\pi(w_r+2s_r)}{4h_s}\right]. \tag{B.20}$$

After the second transformation, the real positive part of the top-half cross-section is

now mapped to a parallel-plate waveguide, with the width and the height

$$W_t = |w_I - w_E| = |A_1|K(k_s), \tag{B.21}$$
$$H_t = |w_P - w_I| = |A_1|K(k_s'), \tag{B.22}$$

where we have used the following relations

$$F(\frac{\pi}{2}, k_s) = K(k_s), \tag{B.23}$$
$$F(\arcsin\frac{1}{k_s}, k_s) = K(k_s) + iK(k_s'), \tag{B.24}$$
$$k_s' = \sqrt{(1 - k_s^2)}. \tag{B.25}$$

After combining the same results from the real negative part of the top-half cross-section, the geometric inductance and capacitance per unit length of the top half of the CPW cross-section are

$$L_l^{g,t} = \frac{\mu_0}{2}\frac{H_t}{W_t} = \frac{\mu_0}{2}\frac{K(k_s')}{K(k_s)}, \tag{B.26}$$
$$C_l^t = 2\varepsilon_0\frac{W_t}{H_t} = 2\varepsilon_0\frac{K(k_s)}{K(k_s')}. \tag{B.27}$$

Therefore, the total geometric inductance and capacitance per unit length of the CPW cross-section in flip-chip geometry, when the CPW line faces a metal ground plane on the top chip, are given as

$$L_l^{g,\mathrm{m}} = \left(1/L_l^{g,t} + 1/L_l^{g,b}\right)^{-1} = \frac{\mu_0}{2}\left[\frac{K(k_s)}{K(k_s')} + \frac{K(k_1)}{K(k_1')}\right]^{-1}, \tag{B.28}$$

$$\begin{aligned} C_l^{\mathrm{m}} &= C_l^t + C_l^b \\ &= 2\varepsilon_0\frac{K(k_s)}{K(k_s')} + 2\varepsilon_0\left[\frac{K(k_1)}{K(k_1')} + (\varepsilon_r - 1)\frac{K(k_2)}{K(k_2')}\right]. \end{aligned} \tag{B.29}$$

B.2.2 Top chip without the ground plane

When the top half cross-section has only the dielectric substrate of the top chip, i.e., no metal ground plane is covering the CPW line, the top chip will only contribute the capacitance C_l^t to the total C_l, due to extra permittivity contribution, while $L_l^{g,t}$ is the same as in the vacuum case.

C_l^t can be calculated in the same way as the bottom half of the cross-section or as the planar CPW line case, in which the capacitance contribution from the vacuum and the substrate of the top chip can be calculated separately and later combined to obtain C_l^t.

The contribution from the substrate is simply by viewing the whole top half filled with the dielectric medium same as the substrate with the relative permittivity ε_r:

$$C_l^{t,\mathrm{sub}} = 2\varepsilon_0\varepsilon_r\frac{K(k_1)}{K(k_1')}. \tag{B.30}$$

However, the contribution from the vacuum is not $2\varepsilon_0(1-\varepsilon_r)\frac{K(k_s)}{K(k_s')}$, but

$$C_l^{t,\mathrm{vac}} = \frac{2\varepsilon_0}{1-\varepsilon_r^{-1}}\frac{K(k_s)}{K(k_s')} \tag{B.31}$$

due to the fact that a lower permittivity medium (vacuum) is closer to the CPW line's center conductor than another medium (top chip's substrate) [92].

And the calculation of C_l^t is not viewing $C_l^{t,\mathrm{sub}}$ and $C_l^{t,\mathrm{vac}}$ in a parallel combination, but in a series combination:

$$\frac{1}{C_l^t} = \frac{1}{C_l^{t,\mathrm{sub}}} + \frac{1}{C_l^{t,\mathrm{vac}}}. \tag{B.32}$$

Therefore, the total geometric inductance and capacitance per unit length of the CPW cross-section in flip-chip geometry, when the CPW line faces the dielectric substrate of the top chip, are given as

$$L_l^{g,\mathrm{d}} = \left(1/L_l^{g,t} + 1/L_l^{g,b}\right)^{-1} = \frac{\mu_0}{4}\frac{K(k_1')}{K(k_1)}, \tag{B.33}$$

$$\begin{aligned} C_l^{\mathrm{d}} &= C_l^t + C_l^b = \left(1/C_l^{t,\mathrm{sub}} + 1/C_l^{t,\mathrm{vac}}\right)^{-1} + C_l^b \\ &= \frac{2\varepsilon_0}{\left(\varepsilon_r\frac{K(k_1)}{K(k_1')}\right)^{-1} + \left(\frac{\varepsilon_r}{\varepsilon_r-1}\frac{K(k_s)}{K(k_s')}\right)^{-1}} \\ &\quad + 2\varepsilon_0\left[\frac{K(k_1)}{K(k_1')} + (\varepsilon_r-1)\frac{K(k_2)}{K(k_2')}\right]. \end{aligned} \tag{B.34}$$

References

[1] R. P. Feynman, "Simulating physics with computers", International Journal of Theoretical Physics **21**, 467–488 (1982).

[2] R. Acharya et al., "Quantum error correction below the surface code threshold", arXiv, 2408.13687 (2024).

[3] D. Castelvecchi, "IBM releases first-ever 1,000-qubit quantum chip", Nature **624**, 238–238 (2023).

[4] F. Arute et al., "Quantum supremacy using a programmable superconducting processor", Nature **574**, 505–510 (2019).

[5] P. Jurcevic et al., "Demonstration of quantum volume 64 on a superconducting quantum computing system", Quantum Science and Technology **6**, 025020 (2021).

[6] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2010).

[7] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, "A quantum engineer's guide to superconducting qubits", Applied Physics Reviews **6**, 021318 (2019).

[8] P. W. Shor, "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer", SIAM Journal on Computing **26**, 1484–1509 (1997).

[9] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems", Communications of the ACM **21**, 120–126 (1978).

[10] L. K. Grover, "A fast quantum mechanical algorithm for database search", in Proceedings of the twenty-eighth annual acm symposium on theory of computing, STOC '96 (1996), pp. 212–219.

[11] A. J. Daley, I. Bloch, C. Kokail, S. Flannigan, N. Pearson, M. Troyer, and P. Zoller, "Practical quantum advantage in quantum simulation", Nature **607**, 667–676 (2022).

[12] B. Cheng et al., "Noisy intermediate-scale quantum computers", Frontiers of Physics **18**, 21308 (2023).

[13] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, "Circuit quantum electrodynamics", Reviews of Modern Physics **93**, 025005 (2021).

[14] Y. Kim et al., "Evidence for the utility of quantum computing before fault tolerance", Nature **618**, 500–505 (2023).

[15] J. V. Damme et al., “Advanced CMOS manufacturing of superconducting qubits on 300 mm wafers”, Nature **634**, 74–79 (2024).
[16] A. Wallraff et al., “Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics”, Nature **431**, 162–167 (2004).
[17] B. Josephson, “Possible new effects in superconductive tunnelling”, Physics Letters **1**, 251–253 (1962).
[18] L. B. Nguyen et al., “Blueprint for a high-performance fluxonium quantum processor”, PRX Quantum **3**, 037001 (2022).
[19] U. Réglade et al., “Quantum control of a cat qubit with bit-flip times exceeding ten seconds”, Nature **629**, 778–783 (2024).
[20] E. Crane et al., “Hybrid Oscillator-Qubit Quantum Processors: Simulating Fermions, Bosons, and Gauge Fields”, arXiv, 2409.03747 (2024).
[21] J. Koch et al., “Charge-insensitive qubit design derived from the Cooper pair box”, Physical Review A **76**, 042319 (2007).
[22] A. Bengtsson et al., “Improved success probability with greater circuit depth for the quantum approximate optimization algorithm”, Physical Review Applied **14**, 1 (2020).
[23] Z. Li, T. Roy, D. Rodríguez Pérez, K.-H. Lee, E. Kapit, and D. I. Schuster, “Autonomous error correction of a single logical qubit using two transmons”, Nature Communications **15**, 1681 (2024).
[24] C. Gidney and M. Ekerå, “How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits”, Quantum **5**, 433 (2021).
[25] P. W. Anderson, “More is different”, Science **177**, 393–396 (1972).
[26] J. Biznárová et al., “Mitigation of interfacial dielectric loss in aluminum-on-silicon superconducting qubits”, npj Quantum Information **10**, 78 (2024).
[27] M. McEwen et al., “Resisting high-energy impact events through gap engineering in superconducting qubit arrays”, arXiv, 2402.15644 (2024).
[28] S. Ling and C. Xing, *Coding theory: a first course* (Cambridge University Press, 2004).
[29] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, “High-threshold and low-overhead fault-tolerant quantum memory”, Nature **627**, 778–782 (2024).
[30] A. D. Córcoles et al., “Challenges and opportunities of near-term quantum computing systems”, Proceedings of the IEEE **108**, 1338–1352 (2020).
[31] A. Javadi-Abhari et al., “Quantum computing with Qiskit”, arXiv, 2405.08810 (2024).
[32] S. Krinner et al., “Engineering cryogenic setups for 100-qubit scale superconducting circuit systems”, EPJ Quantum Technology **6**, 2 (2019).
[33] M. McEwen et al., “Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits”, Nature Physics **18**, 107–111 (2022).
[34] R. Tummala, *Fundamentals of device and systems packaging: technologies and applications, second edition* (McGraw Hill LLC, 2019).
[35] J. Lau, *Chiplet design and heterogeneous integration packaging* (Springer Nature Singapore, 2023).

[36] J. H. Lau, "Recent advances and trends in advanced packaging", IEEE Transactions on Components, Packaging and Manufacturing Technology **12**, 228–252 (2022).
[37] R. Schaller, "Moore's law: past, present and future", IEEE Spectrum **34**, 52–59 (1997).
[38] D. Lu and C. Wong, *Materials for advanced packaging* (Springer International Publishing, 2016).
[39] A. Chen and R. Lo, *Semiconductor packaging: materials interaction and reliability* (CRC Press, 2016).
[40] N. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective* (Addison Wesley, 2011).
[41] C. Müller, J. H. Cole, and J. Lisenfeld, "Towards understanding two-level-systems in amorphous solids: Insights from quantum circuits", Reports on Progress in Physics **82**, 124501 (2019).
[42] Z. Chen et al., "Fabrication and characterization of aluminum airbridges for superconducting microwave circuits", Applied Physics Letters **104**, 052602 (2014).
[43] S. Krinner et al., "Realizing repeated quantum error correction in a distance-three surface code", Nature **605**, 669–674 (2022).
[44] B. Foxen et al., "Qubit compatible superconducting interconnects", Quantum Science and Technology **3**, 014005 (2018).
[45] P. A. Spring et al., "High coherence and low cross-talk in a tileable 3D integrated superconducting circuit architecture", Science Advances **8**, 1–18 (2022).
[46] J. H. Béjanin et al., "Three-Dimensional Wiring for Extensible Quantum Computing: The Quantum Socket", Physical Review Applied **6**, 044010 (2016).
[47] D. R. W. Yost et al., "Solid-state qubits integrated with superconducting through-silicon vias", npj Quantum Information **6**, 59 (2020).
[48] B. El-Kareh and L. Hutter, *Silicon analog components: device design, process integration, characterization, and reliability* (Springer International Publishing, 2019).
[49] C. Alpert, D. Mehta, and S. Sapatnekar, *Handbook of algorithms for physical design automation* (Taylor & Francis, 2008).
[50] C. Neill et al., "A blueprint for demonstrating quantum supremacy with superconducting qubits", Science **360**, 195–199 (2018).
[51] J. J. Burnett, A. Bengtsson, M. Scigliuzzo, D. Niepce, M. Kudra, P. Delsing, and J. Bylander, "Decoherence benchmarking of superconducting qubits", npj Quantum Information **5**, 54 (2019).
[52] J. I.-J. Wang et al., "Hexagonal boron nitride as a low-loss dielectric for superconducting quantum circuits and qubits", Nature Materials **21**, 398–403 (2022).
[53] Y. Sun, N. B. Agostini, S. Dong, and D. R. Kaeli, "Summarizing CPU and GPU design trends with product data", arXiv, 1911.11313 (2019).
[54] N. Takeuchi, T. Yamae, T. Yamashita, T. Yamamoto, and N. Yoshikawa, "Microwave-multiplexed qubit controller using adiabatic superconductor logic", npj Quantum Information **10**, 53 (2024).
[55] Z. Bao et al., "A cryogenic on-chip microwave pulse generator for large-scale superconducting quantum computing", Nature Communications **15**, 5958 (2024).

[56] D. Rosenberg et al., “3D integrated superconducting qubits”, npj Quantum Information **3**, 42 (2017).

[57] M. Vahidpour et al., “Superconducting through-silicon vias for quantum integrated circuits”, arXiv, 1708.02226 (2017).

[58] T. M. Hazard et al., “Characterization of superconducting through-silicon vias as capacitive elements in quantum circuits”, Applied Physics Letters **123**, 154004 (2023).

[59] L. Chayanun et al., “Characterization of process-related interfacial dielectric loss in aluminum-on-silicon by resonator microwave measurements, materials analysis, and imaging”, APL Quantum **1**, 026115 (2024).

[60] A. Osman et al., “Simplified Josephson-junction fabrication process for reproducibly high-performance superconducting qubits”, Applied Physics Letters **118**, 064002 (2021).

[61] D. C. McKay, S. Filipp, A. Mezzacapo, E. Magesan, J. M. Chow, and J. M. Gambetta, “Universal Gate for Fixed-Frequency Qubits via a Tunable Bus”, Physical Review Applied **6**, 064007 (2016).

[62] M. Roth, M. Ganzhorn, N. Moll, S. Filipp, G. Salis, and S. Schmidt, “Analysis of a parametrically driven exchange-type gate and a two-photon excitation gate between superconducting qubits”, Physical Review A **96**, 062323 (2017).

[63] E. Magesan, J. M. Gambetta, and J. Emerson, “Scalable and robust randomized benchmarking of quantum processes”, Physical Review Letters **106**, 180504 (2011).

[64] E. M. Purcell, “Spontaneous Emission Probabilities at Radio Frequencies”, in *Confined electrons and photons* (Springer New York, NY, 1995), pp. 839–839.

[65] M. D. Reed, B. R. Johnson, A. A. Houck, L. DiCarlo, J. M. Chow, D. I. Schuster, L. Frunzio, and R. J. Schoelkopf, “Fast reset and suppressing spontaneous emission of a superconducting qubit”, Applied Physics Letters **96**, 203110 (2010).

[66] J. Heinsoo et al., “Rapid high-fidelity multiplexed readout of superconducting qubits”, Physical Review Applied **10**, 034040 (2018).

[67] E. Jeffrey et al., “Fast Accurate State Measurement with Superconducting Qubits”, Physical Review Letters **112**, 190504 (2014).

[68] D. Pozar, *Microwave engineering 4th edition* (Wiley, 2012).

[69] M. Field et al., “Modular superconducting-qubit architecture with a multichip tunable coupler”, Physical Review Applied **21**, 054063 (2024).

[70] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, “Surface codes: towards practical large-scale quantum computation”, Physical Review A **86**, 032324 (2012).

[71] R. Acharya et al., “Suppressing quantum errors by scaling a surface code logical qubit”, Nature **614**, 676–681 (2023).

[72] X. Zhang et al., “Digital quantum simulation of Floquet symmetry-protected topological phases”, Nature **607**, 468–473 (2022).

[73] A. Osman et al., “Mitigation of frequency collisions in superconducting quantum processors”, Physical Review Research **5**, 043001 (2023).

[74] B. Li, T. Calarco, and F. Motzoi, “Experimental error suppression in Cross-Resonance gates via multi-derivative pulse shaping”, npj Quantum Information **10**, 66 (2024).

[75] P. V. Klimov et al., "Optimizing quantum gates towards the scale of logical qubits", Nature Communications **15**, 2442 (2024).
[76] J. Chu and F. Yan, "Coupler-Assisted Controlled-Phase Gate with Enhanced Adiabaticity", Physical Review Applied **16**, 054020 (2021).
[77] X. Ni, H.-H. Zhao, L. Wang, F. Wu, and J. Chen, "Integrating quantum processor device and control optimization in a gradient-based framework", npj Quantum Information **8**, 106 (2022).
[78] T. Abad, J. Fernández-Pendás, A. F. Kockum, and G. Johansson, "Universal fidelity reduction of quantum operations from weak dissipation", Physical Review Letters **129**, 150504 (2022).
[79] S. P. Fors, J. Fernández-Pendás, and A. F. Kockum, "Comprehensive explanation of ZZ coupling in superconducting qubits", arXiv, 2408.15402 (2024).
[80] D. Sank, A. Opremcak, A. Bengtsson, M. Khezri, Z. Chen, O. Naaman, and A. Korotkov, "System characterization of dispersive readout in superconducting qubits", arXiv, 2402.00413 (2024).
[81] A. Osman, *Scaling superconducting quantum processors: coherence, frequency targeting and crosstalk*, Doktorsavhandlingar vid Chalmers tekniska högskola. Ny serie: 5597 (Chalmers, 2024).
[82] *Ansys® Electronic Desktop, HFSS, Release 2020 R1.*
[83] P. A. Shah et al., *qiskit-community/qiskit-metal: Qiskit Metal*, version 0.1.5a1, June 2023.
[84] *Mentor® L-edit*, version 2020.1 Update 4.
[85] J. M. Martinis, "Qubit metrology for building a fault-tolerant quantum computer", npj Quantum Information **1**, 15005 (2015).
[86] J. Anders et al., "CMOS integrated circuits for the quantum information sciences", IEEE Transactions on Quantum Engineering **4**, 5100230 (2023).
[87] H. Mamin et al., "Merged-element transmons: design and qubit performance", Physical Review Applied **16**, 024023 (2021).
[88] R. Schinzinger and P. A. A. Laura, *Conformal mapping: methods and applications* (Dover Publications, 2003).
[89] C. Wen, "Coplanar waveguide: a surface strip transmission line suitable for non-reciprocal gyromagnetic device applications", IEEE Transactions on Microwave Theory and Techniques **17**, 1087–1090 (1969).
[90] G. Ghione and C. Naldi, "Coplanar waveguides for MMIC applications: effect of upper shielding, conductor backing, finite-extent ground planes, and line-to-line coupling", IEEE Transactions on Microwave Theory and Techniques **35**, 260–267 (1987).
[91] S. Gevorgian, L. Linner, and E. Kollberg, "CAD models for shielded multilayered CPW", IEEE Transactions on Microwave Theory and Techniques **43**, 772–779 (1995).
[92] G. Ghione and M. Goano, "Revisiting the partial-capacitance approach to the analysis of coplanar transmission lines on multilayered substrates", IEEE Transactions on Microwave Theory and Techniques **51**, 2007–2014 (2003).
[93] J. Bardeen, "Two-fluid model of superconductivity", Physical Review Letters **1**, 399–400 (1958).

[94] R. Meservey and P. M. Tedrow, "Measurements of the kinetic inductance of superconducting linear structures", Journal of Applied Physics **40**, 2028–2034 (1969).

[95] K. Watanabe, K. Yoshida, T. Aoki, and S. Kohjiro, "Kinetic Inductance of Superconducting Coplanar Waveguides", Japanese Journal of Applied Physics **33**, 5708–5712 (1994).

[96] O. G. Vendik, I. B. Vendik, and D. I. Kaparkov, "Emperical model of the microwave properties of high-temperature superconductors", IEEE Transactions on Microwave Theory and Techniques **46**, 469–478 (1998).

[97] D. Niepce, J. J. Burnett, M. G. Latorre, and J. Bylander, "Geometric scaling of two-level-system loss in superconducting resonators", Superconductor Science and Technology **33**, 025013 (2020).

[98] P. B. Miller, "Penetration depth in impure superconductors", Physical Review **113**, 1209–1212 (1959).

[99] T. P. Orlando and K. A. Delin, *Foundations of applied superconductivity* (Addison-Wesley, 1991).

[100] R. J. C. Charles P. Poole Jr., Horacio A. Farach, *Superconductivity* (Elsevier, 2007).

[101] I. Besedin and A. P. Menushenkov, "Quality factor of a transmission line coupled coplanar waveguide resonator", EPJ Quantum Technology **5**, 2 (2018).